Microsoft

Accelerating Persistent Neural Networks at Datacenter Scale

Eric Chung, Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Adrian Caulfield, Todd Massengil, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, Christian Boehn, Oren Firestein, Alessandro Forin, Kang Su Gatlin, Mahdi Ghandi, Stephen Heil, Kyle Holohan, Tamas Juhasz, Ratna Kumar Kovvuri, Sitaram Lanka, Friedel van Megen, Dima Mukhortov, Prerak Patel, Steve Reinhardt, Adam Sapek, Raja Seera, Balaji Sridharan, Lisa Woods, Phillip Yi-Xiao, Ritchie Zhao, Doug Burger

The Rise of Deep Learning in ML

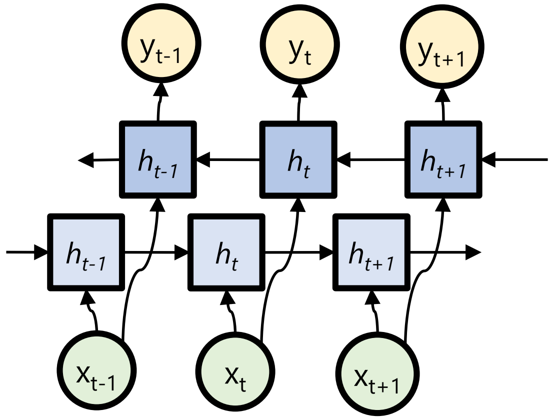
Deep neural networks have enabled major advances in machine learning and AI

- Computer vision
- Language translation
- Speech recognition
- Question answering
- And more…

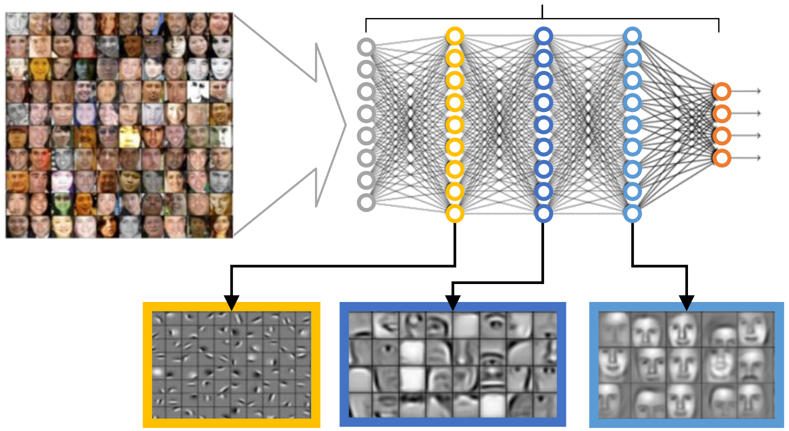
Problem: DNNs are challenging to serve and deploy in large-scale online services

- Heavily constrained by latency, cost, and power
- Size and complexity of DNNs outpacing growth of commodity CPUs

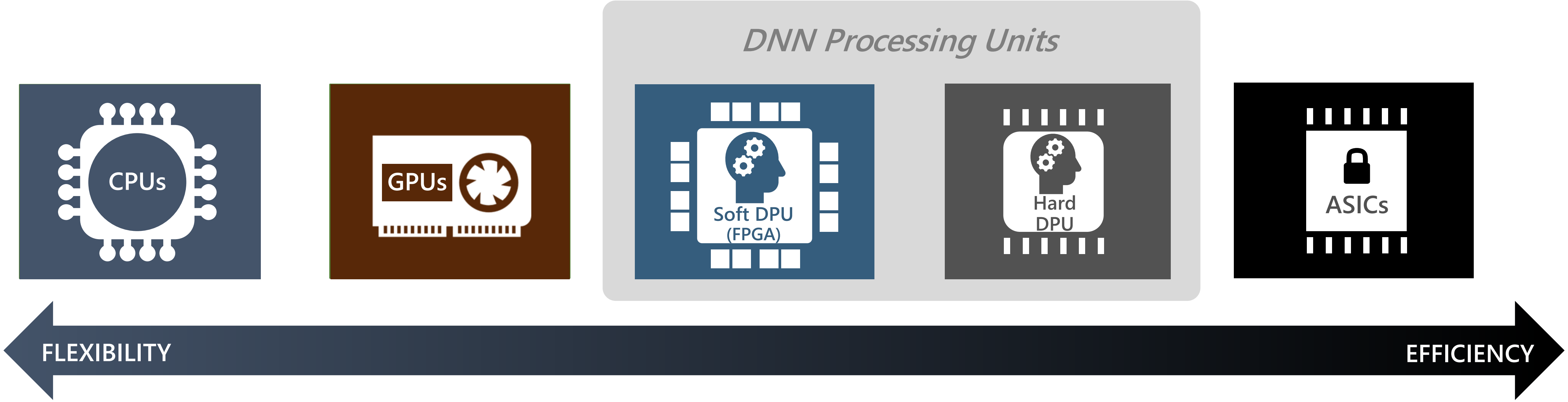
Recurrent Neural Networks



Convolutional Neural Networks



Silicon alternatives for DNNs



MS BrainWave
Baidu SDA
Deephi Tech
ESE
Teradeep
Etc.

Cerebras
Google TPU
Graphcore
Groq
Intel Nervana
Movidius
Wave Computing
Etc.

The power of Deep Learning on FPGA

Performance

Excellent inference performance at low batch sizes
Ultra-low latency serving on modern DNNs
 >10X lower than CPUs and GPUs
Scale to many FPGAs in single DNN service

Flexibility

FPGAs ideal for adapting to rapidly evolving ML
CNNs, LSTMs, MLPs, reinforcement learning, feature extraction, decision trees, etc.
Inference-optimized numerical precision
Exploit sparsity, deep compression for larger, faster models

Scale

Microsoft has the world's largest cloud investment in FPGAs
Multiple Exa-Ops of aggregate AI capacity
BrainWave runs on Microsoft's scale infrastructure

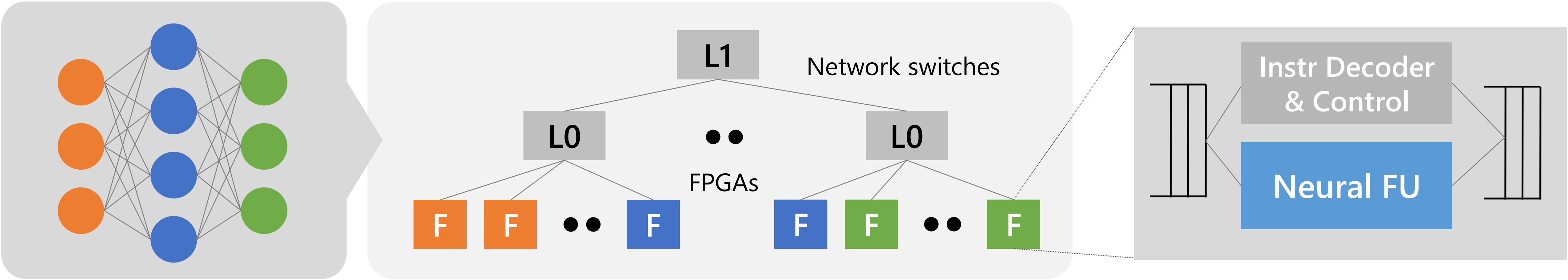
Project BrainWave

A Scalable FPGA-powered DNN Serving Platform

Fast: ultra-low latency, high-throughput serving of DNN models at low batch sizes

Flexible: adaptive numerical precision and custom operators

Friendly: turnkey deployment of CNTK/Caffe/TF/etc

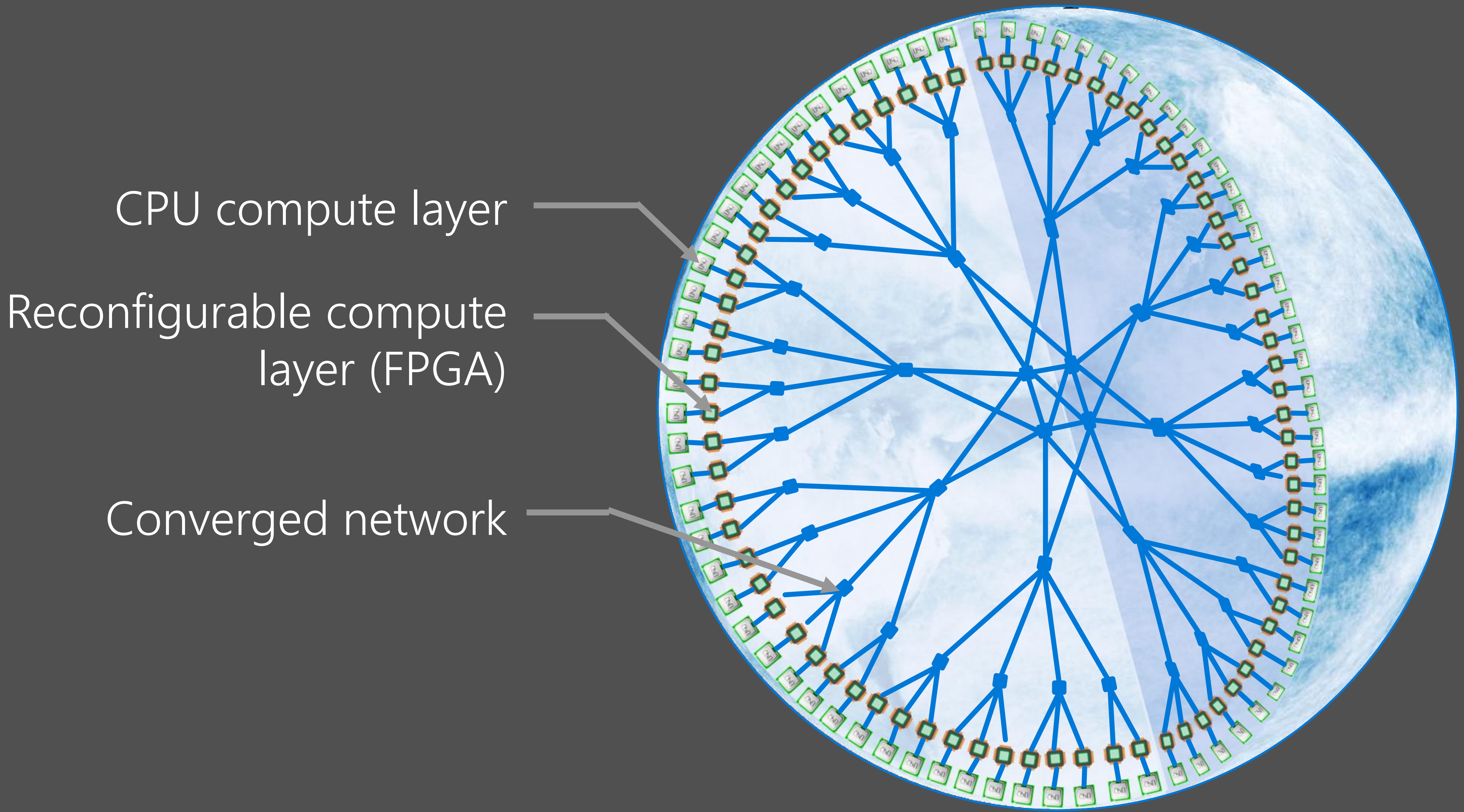


Pretrained DNN Model in CNTK, etc.

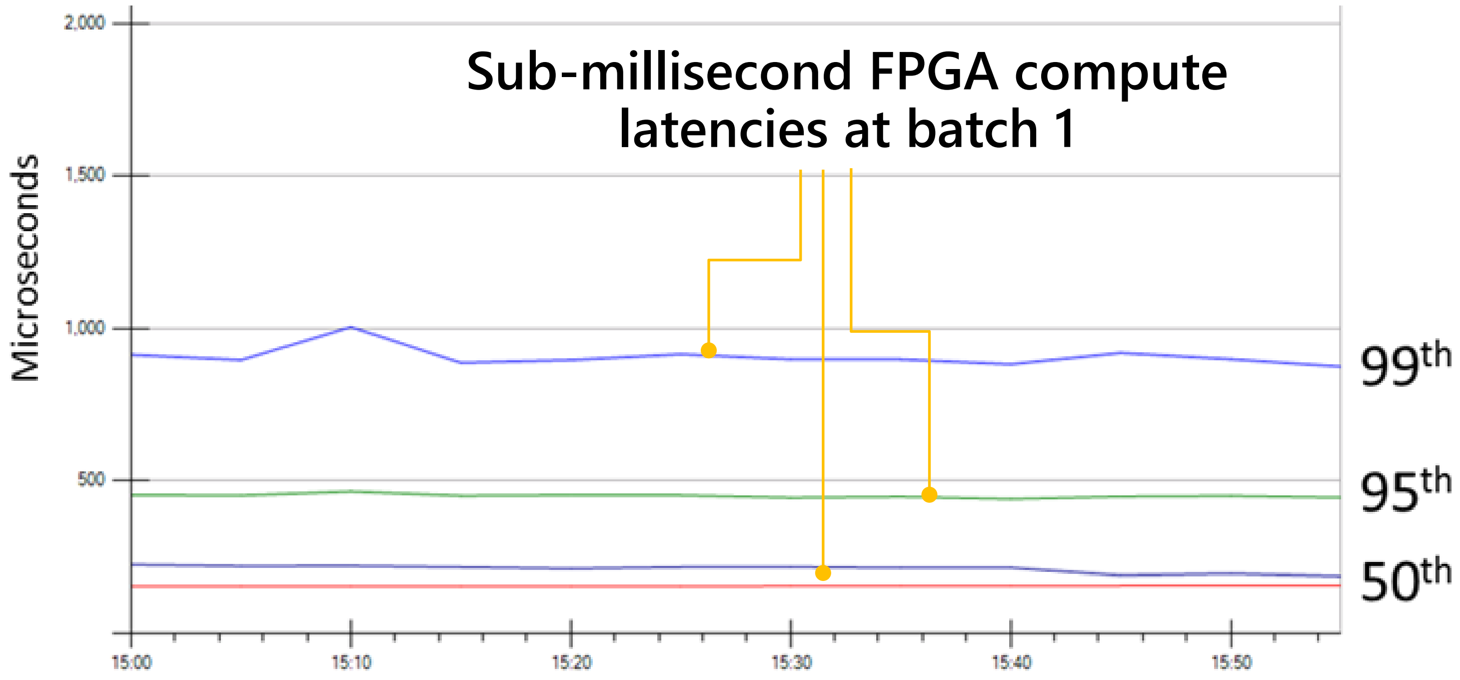
Scalable DNN Hardware Microservice

BrainWave Soft DPU

Runs on a Configurable Cloud at Massive Scale



Deployed in Production Datacenters

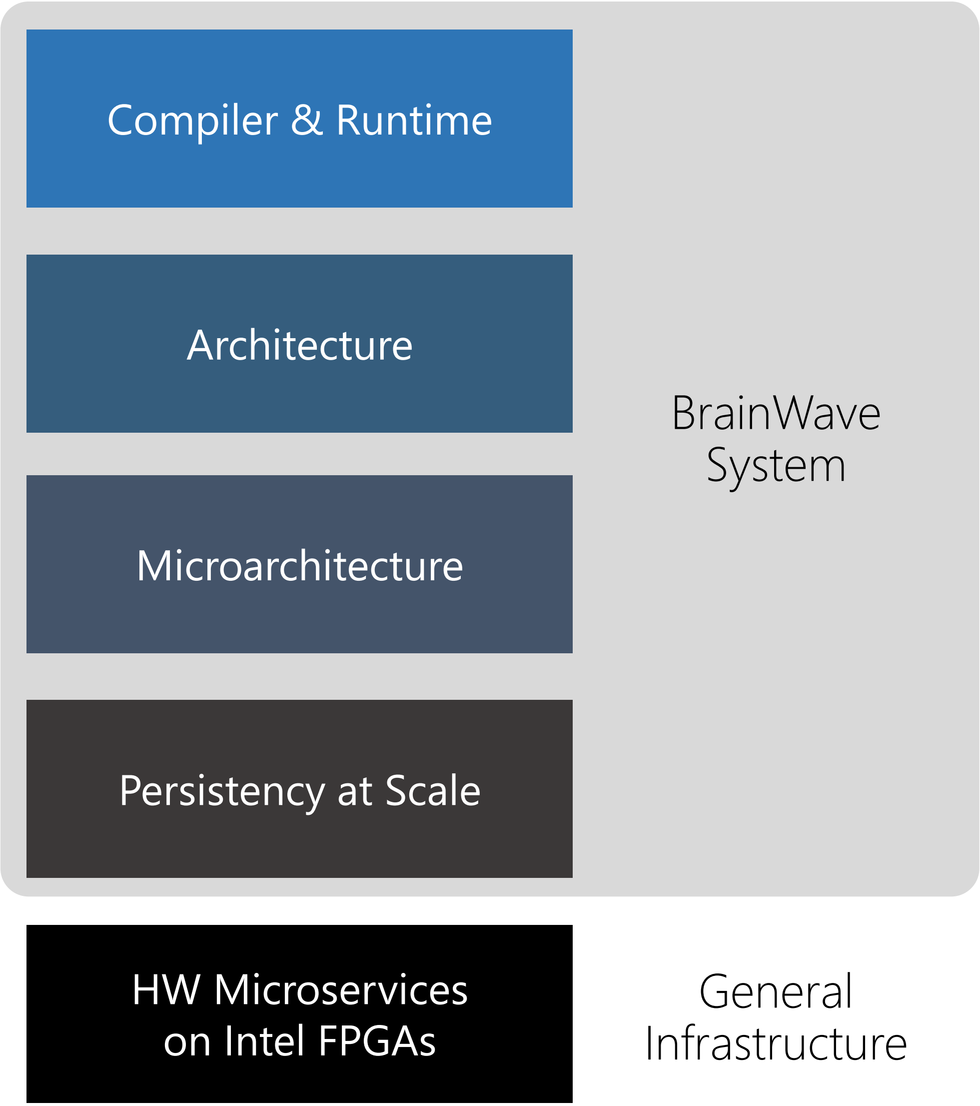


Deployment of LSTM-based NLP model (tens of millions of parameters)

Takes tens of milliseconds to serve on well-tuned CPU implementations

Tail latencies in BrainWave-powered DNN models appear negligible in E2E software pipelines

How It Works: The BrainWave Stack



How It Works: The BrainWave Stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft DPUs
Architecture	
Microarchitecture	
Persistency at Scale	
HW Microservices on Intel FPGAs	

How It Works: The BrainWave Stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft DPUs
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	
Persistency at Scale	
HW Microservices on Intel FPGAs	

How It Works: The BrainWave Stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft DPUs
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	BrainWave Soft DPU microarchitecture Highly optimized for narrow precision and low batch
Persistency at Scale	
HW Microservices on Intel FPGAs	

How It Works: The BrainWave Stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft DPUs
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	BrainWave Soft DPU microarchitecture Highly optimized for narrow precision and low batch
Persistency at Scale	Persist model parameters entirely in FPGA on-chip memories Support large models by scaling across many FPGAs
HW Microservices on Intel FPGAs	

How It Works: The BrainWave Stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft DPUs
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	BrainWave Soft DPU microarchitecture Highly optimized for narrow precision and low batch
Persistency at Scale	Persist model parameters entirely in FPGA on-chip memories Support large models by scaling across many FPGAs
HW Microservices on Intel FPGAs	Intel FPGAs deployed at scale with HW microservices [MICRO'16]

The BrainWave Stack

Compiler & Runtime

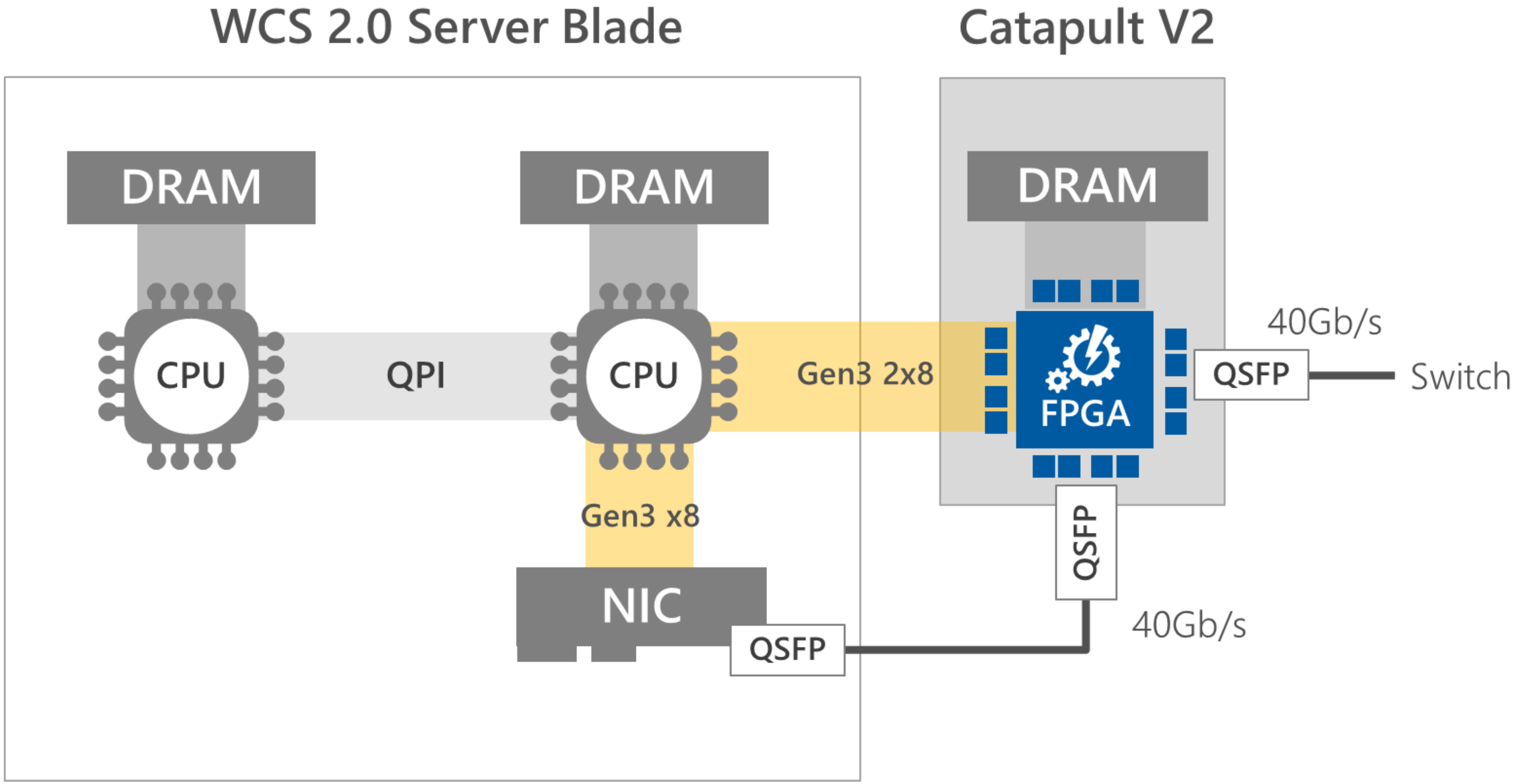
Architecture

Microarchitecture

Persistency at Scale

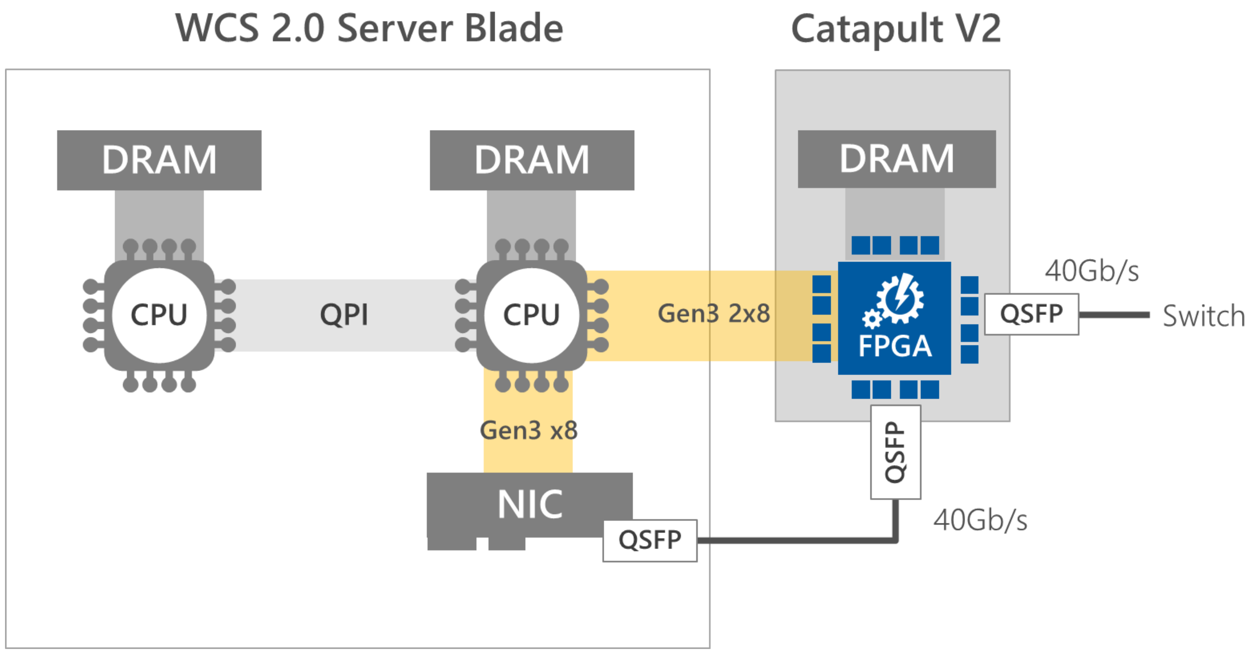
HW Microservices
on Intel FPGAs

FPGAs Are Deployed in MSFT Servers Worldwide



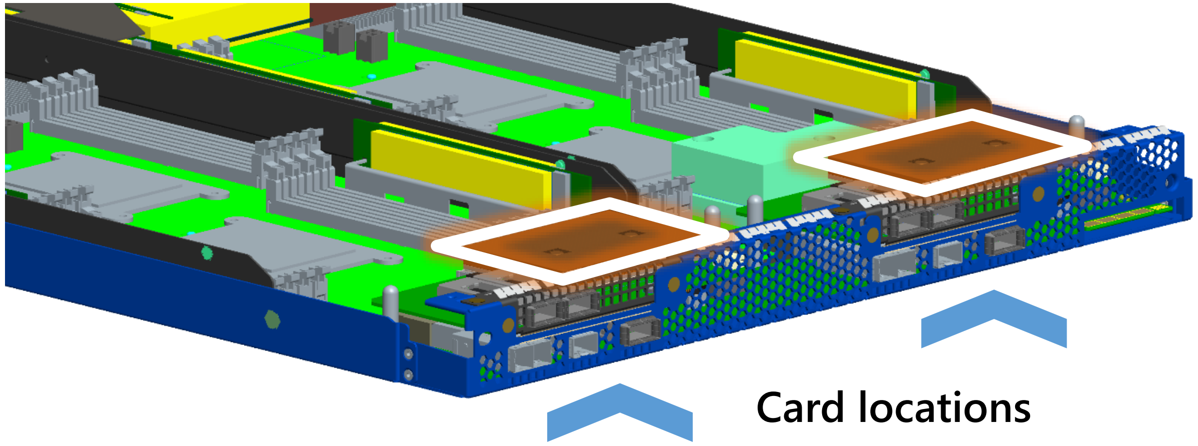
[ISCA'14, HotChips'14, MICRO'16]

FPGAs Are Deployed in MSFT Servers Worldwide



Catapult v2 Mezzanine card

WCS Gen4.1 Blade with NIC and Catapult FPGA

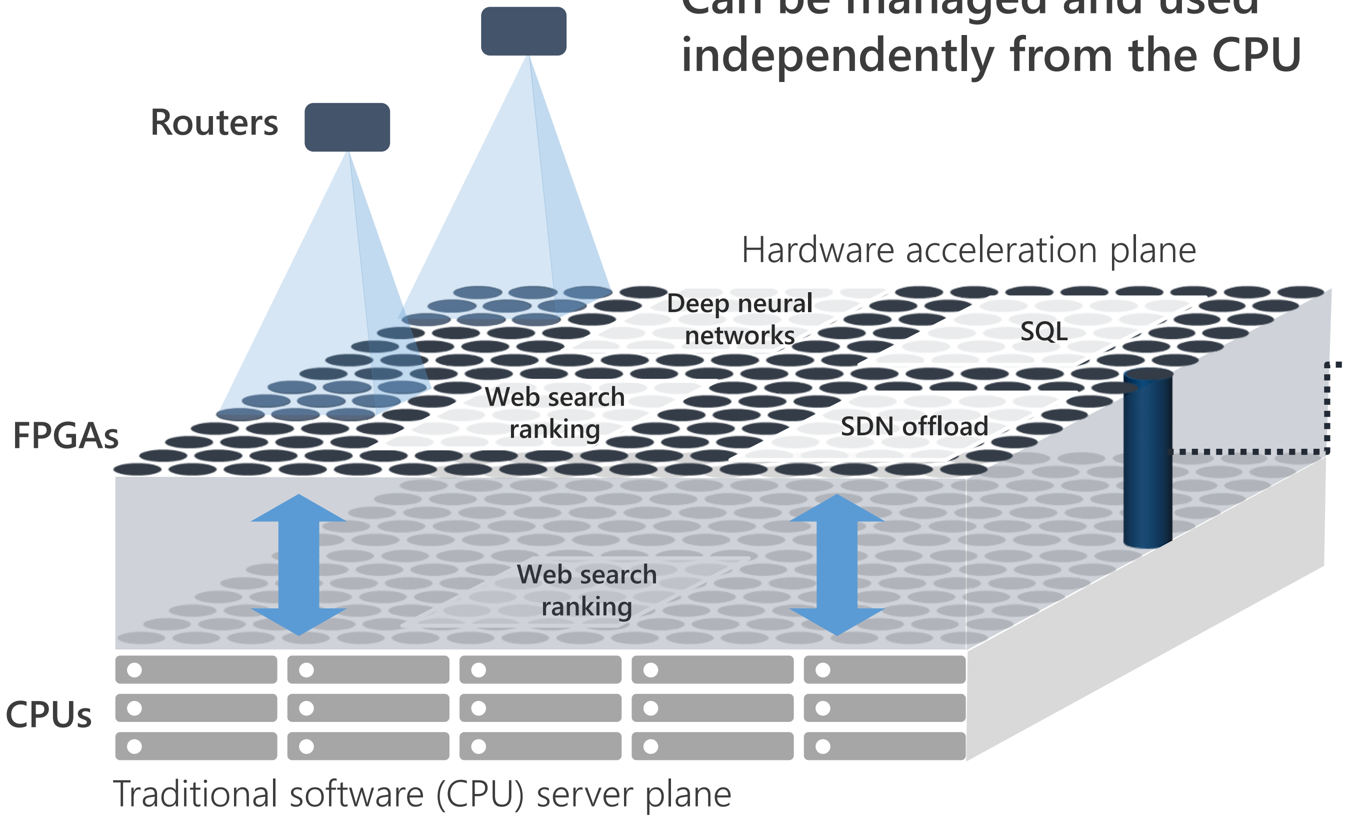


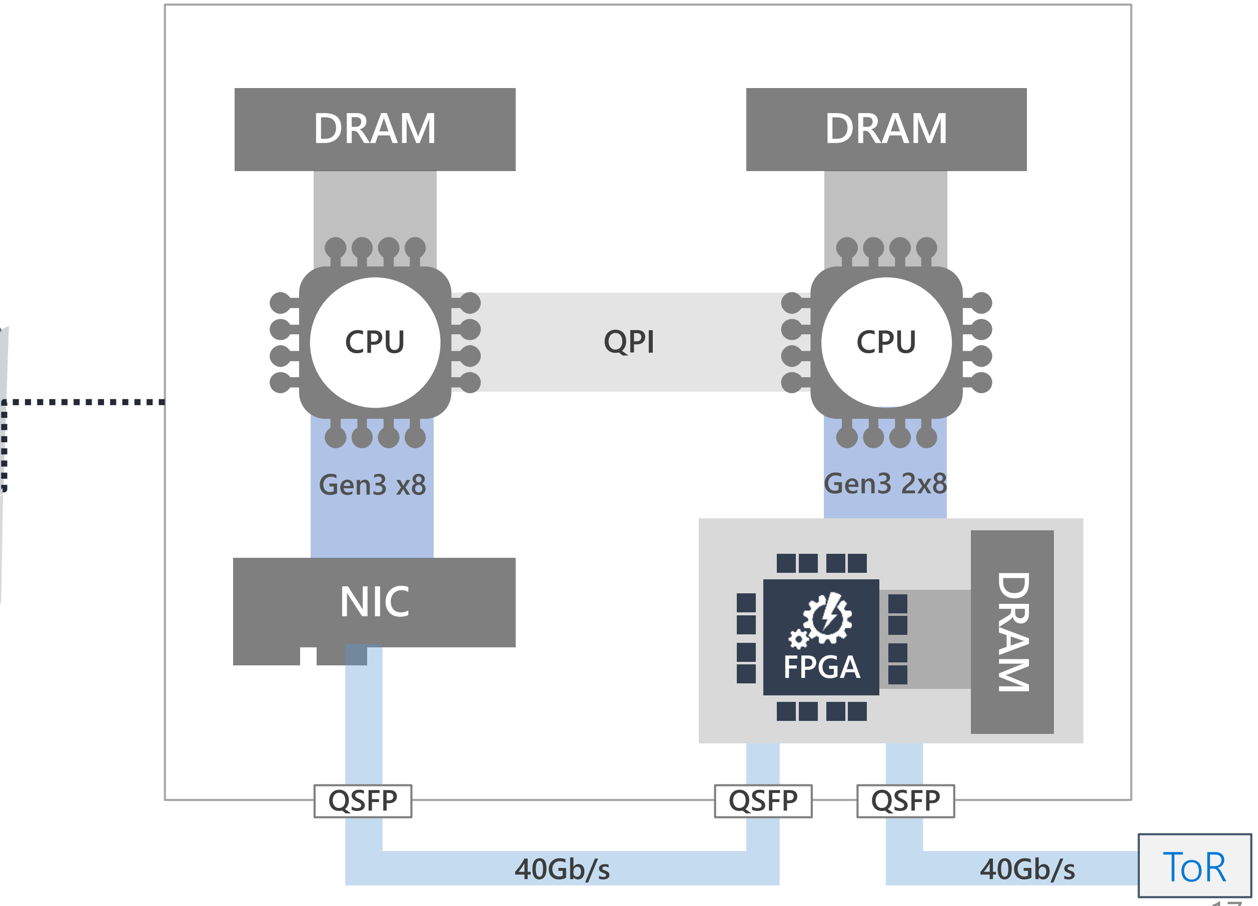
[ISCA'14, HotChips'14, MICRO'16]

Hardware Microservices on FPGAs [MICRO'16]

Interconnected FPGAs form a separate plane of computation

Can be managed and used independently from the CPU





The BrainWave Stack

Compiler & Runtime

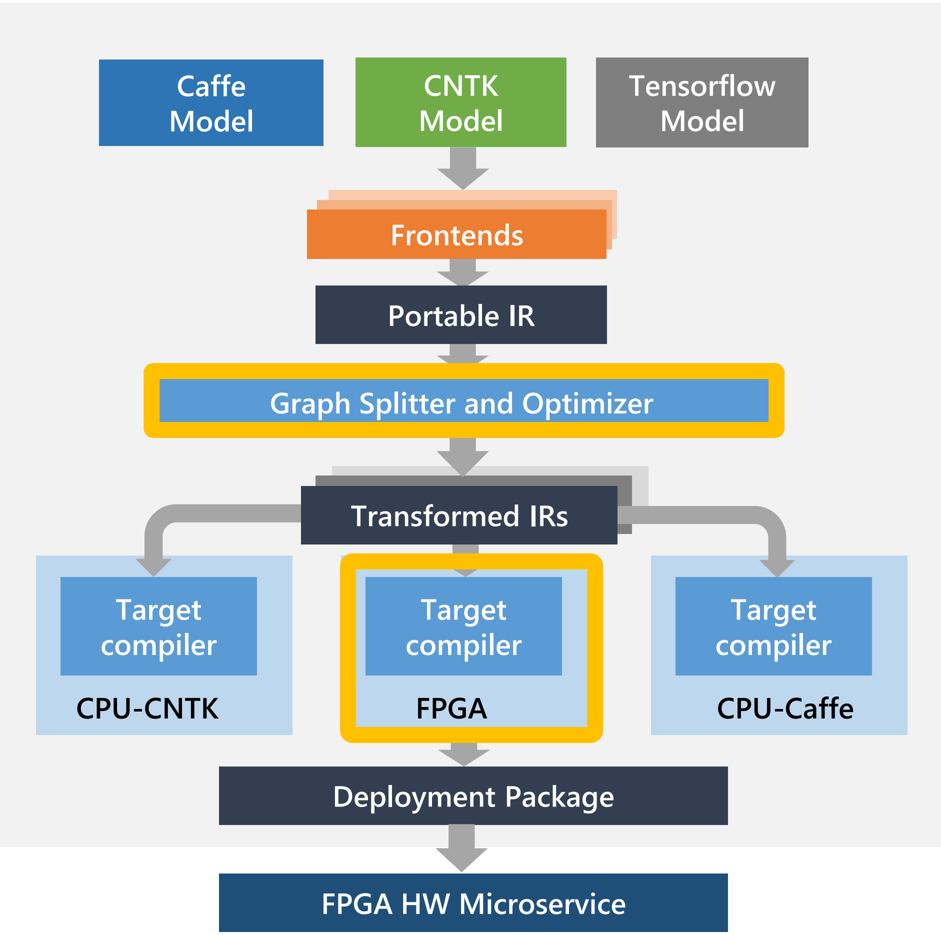
Architecture

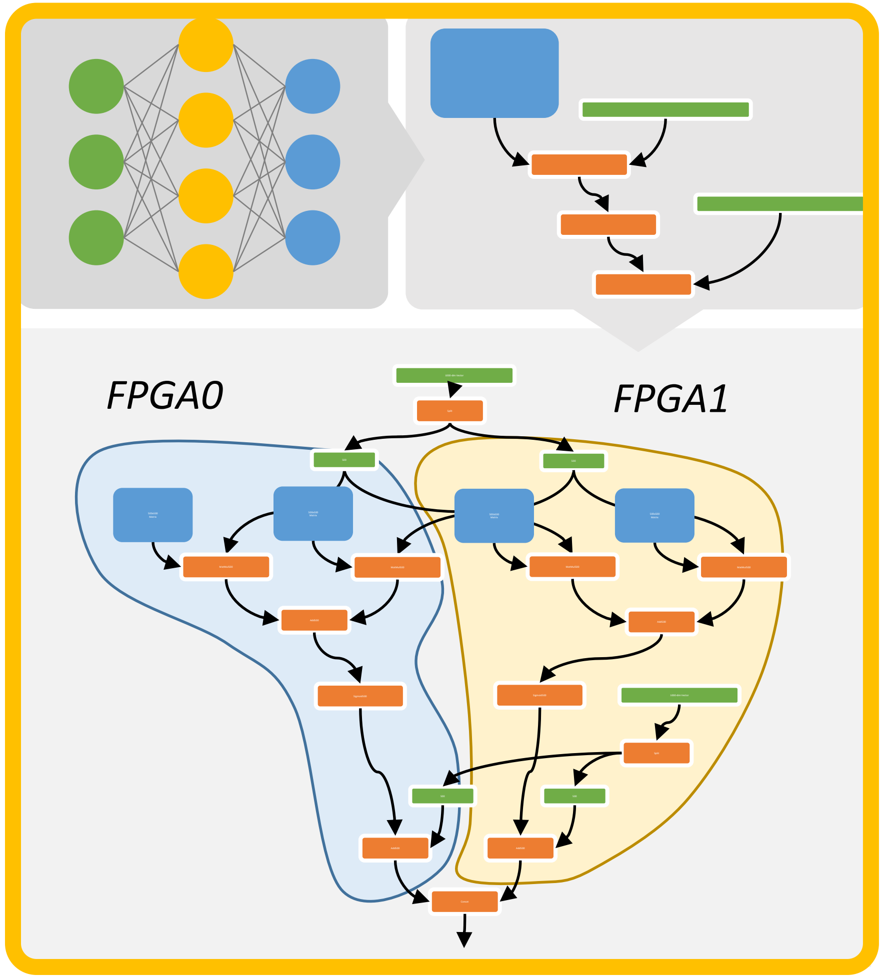
Microarchitecture

Persistency at Scale

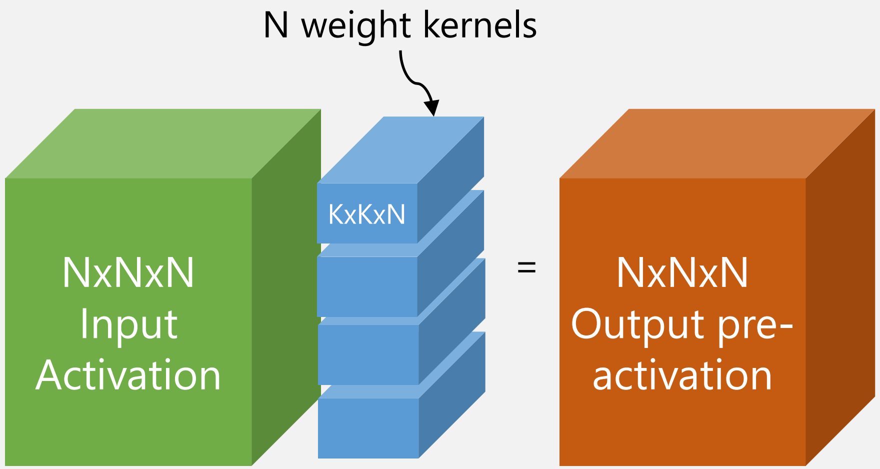
HW Microservices
on Intel FPGAs

BrainWave Compiler & Runtime

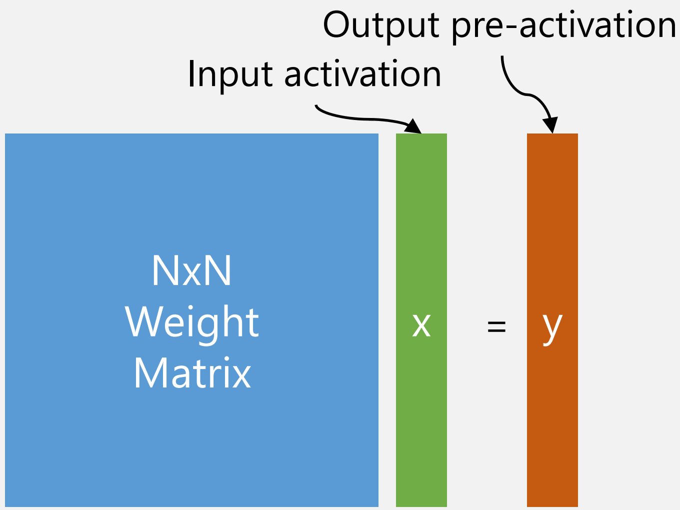




Common Scenarios



$O(N^3)$ data
$O(N^4K^2)$ compute



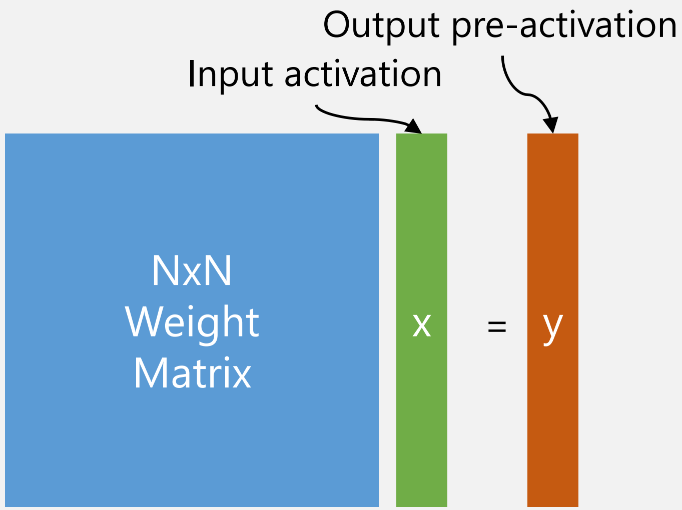
$O(N^2)$ data
$O(N^2)$ compute

Convolutional Neural Network (CNN)
High Compute-to-Data Ratio

MLPs, LSTMs, GRUs
Low compute-to-data ratio

Common Scenarios

$O(N^3)$ data
$O(N^4K^2)$ compute

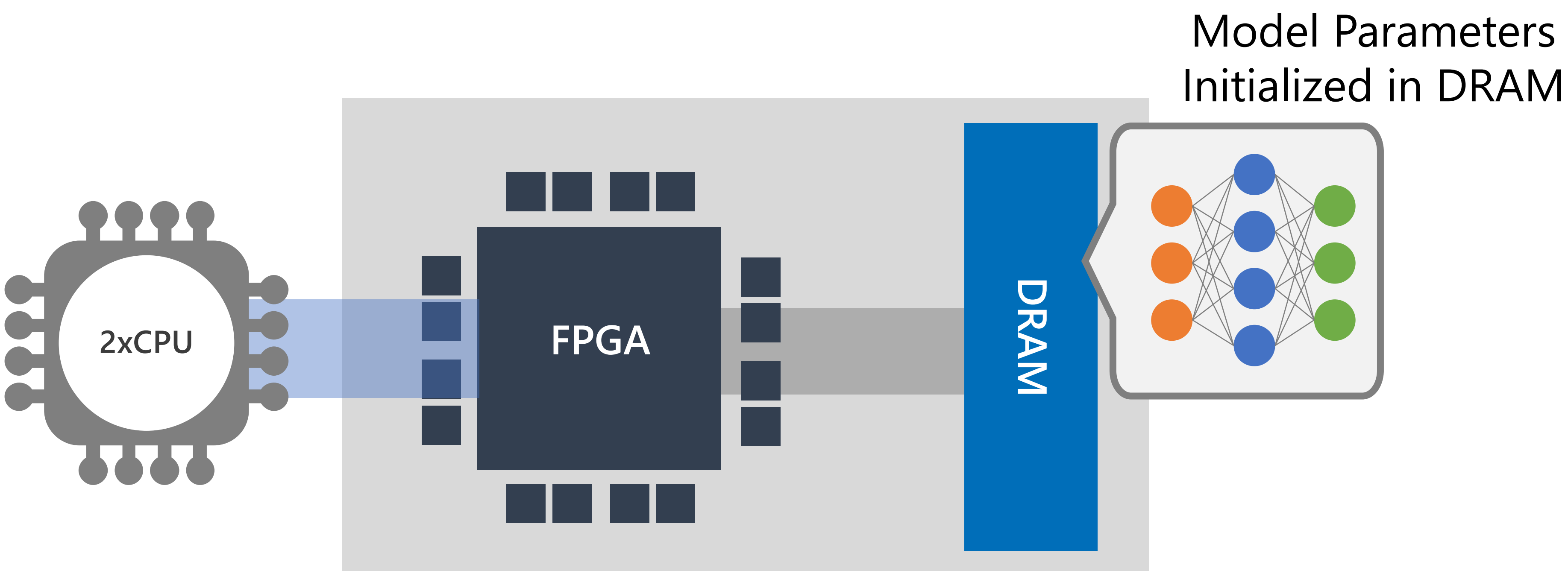


$O(N^2)$ data
$O(N^2)$ compute

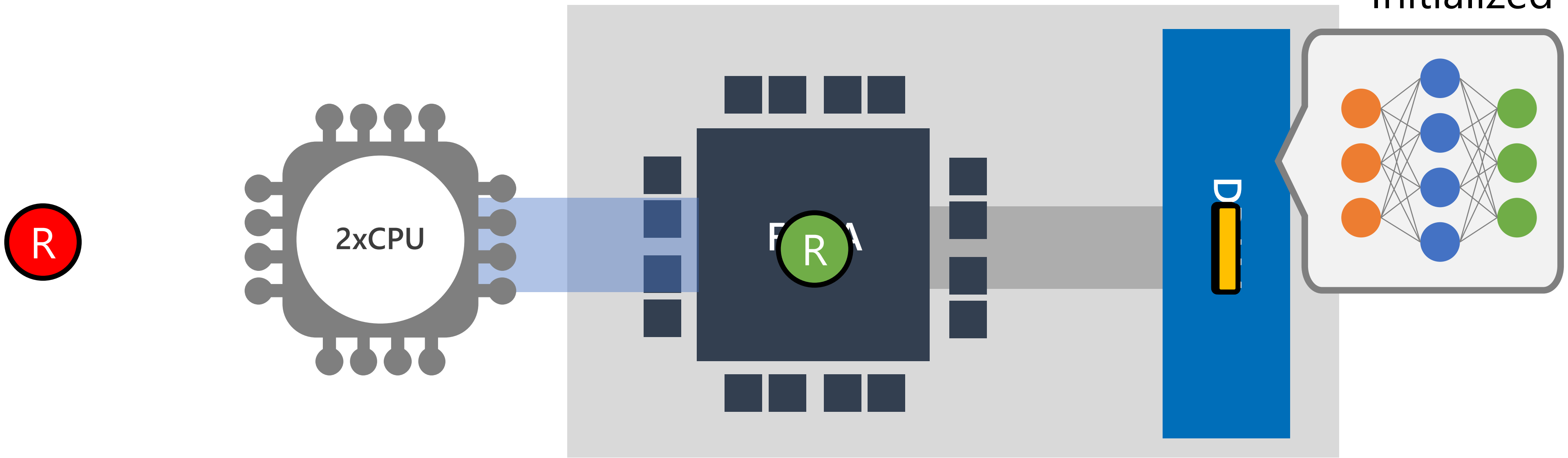
Convolutional Neural Network (CNN)
High Compute-to-Data Ratio

MLPs, LSTMs, GRUs
Low compute-to-data ratio

Conventional Acceleration Approach: Local Offload and Streaming

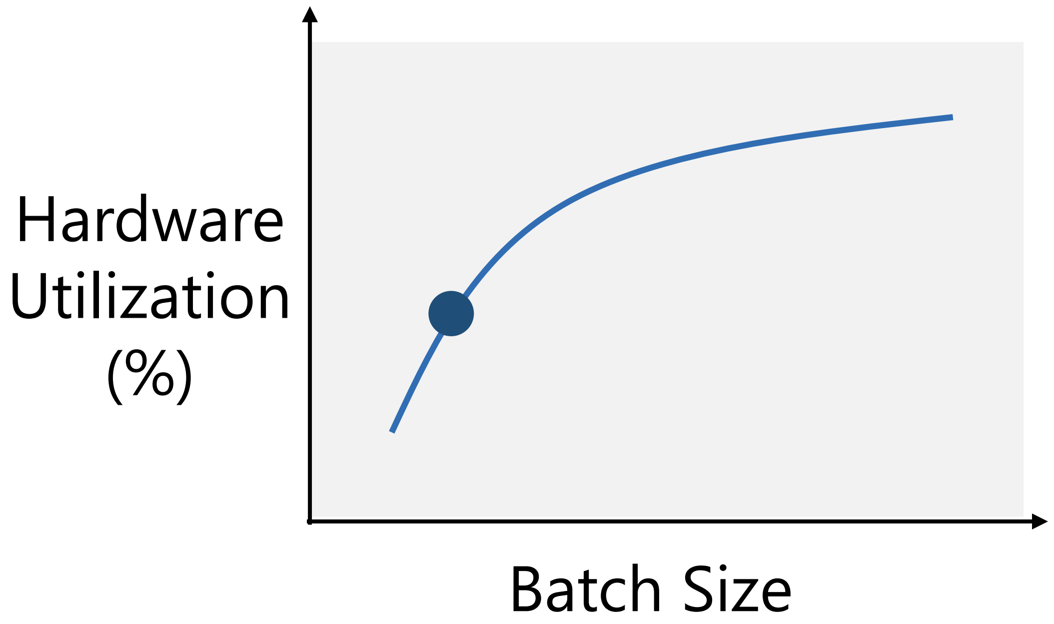


Conventional Acceleration Approach: Local Offload and Streaming

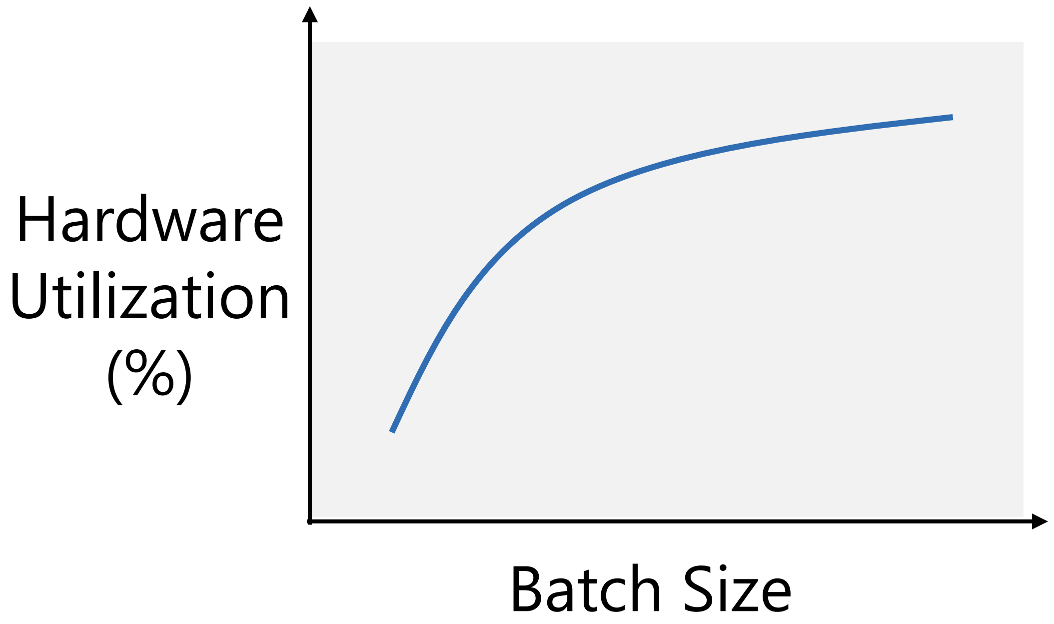


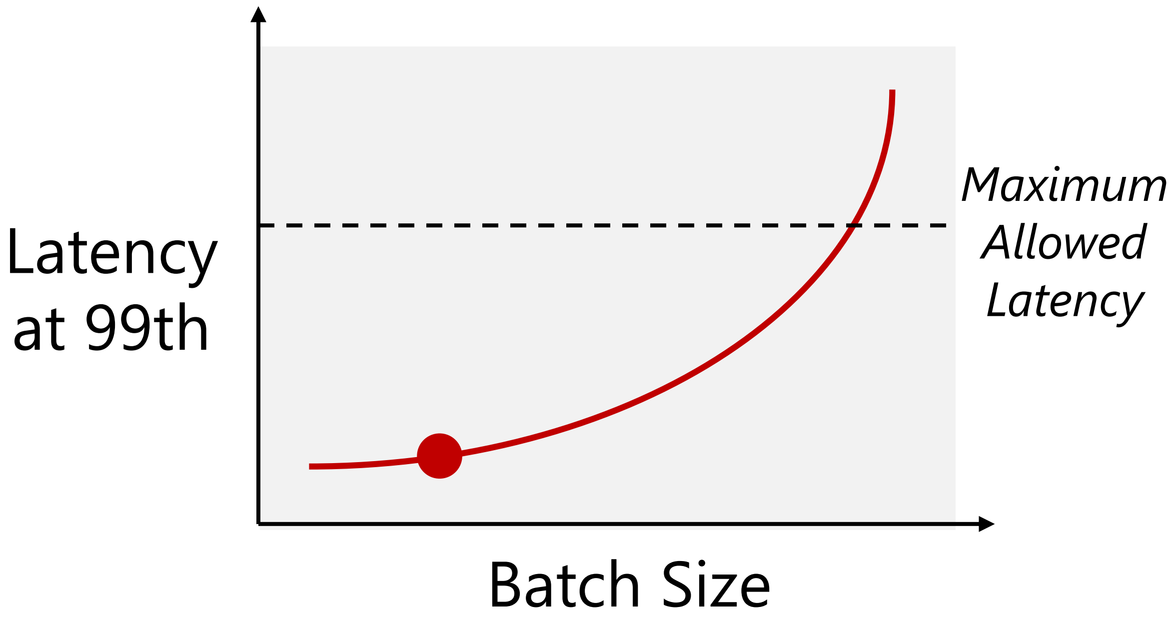
For memory-intensive DNNs with low compute-to-data ratios (e.g., LSTM), HW utilization limited by off-chip DRAM bandwidth

Improving HW utilization with batching



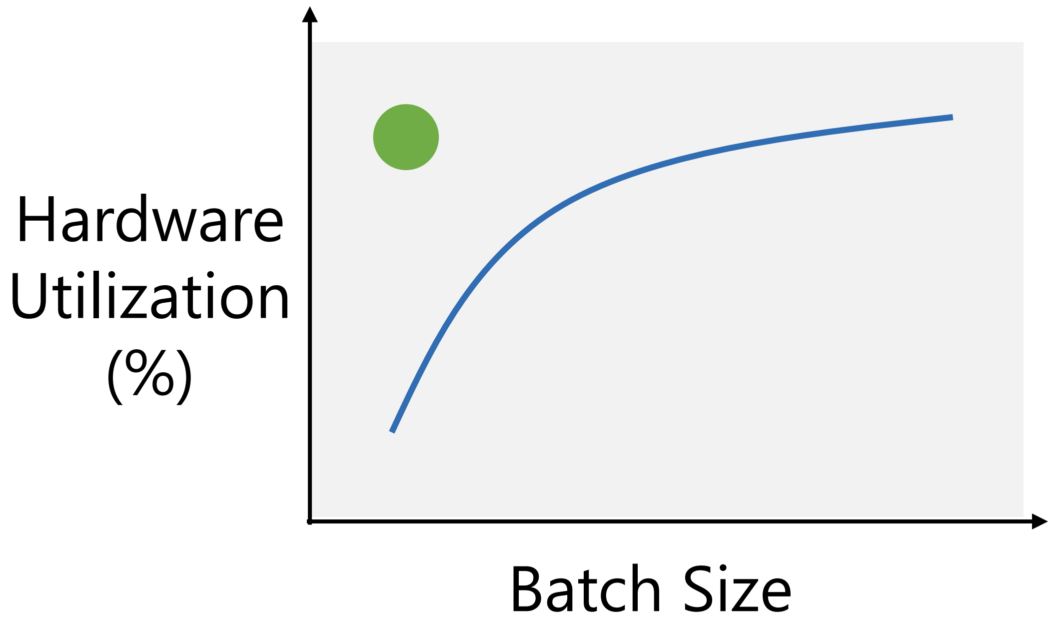
Improving HW utilization with batching

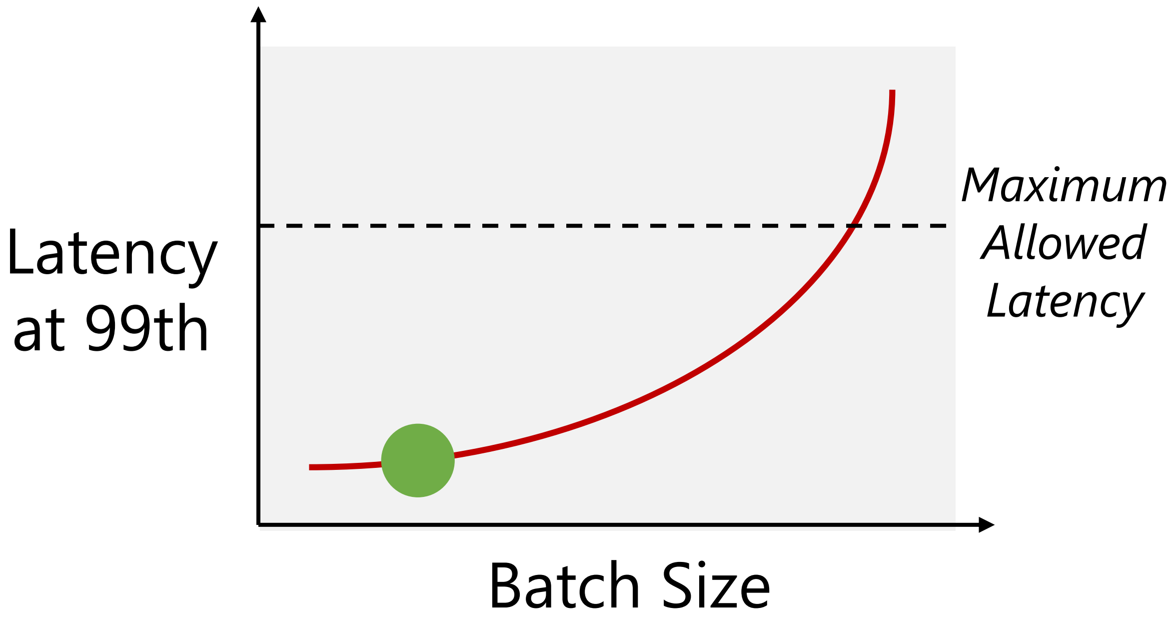




Batching improves HW utilization but increases latency

Improving HW utilization with batching

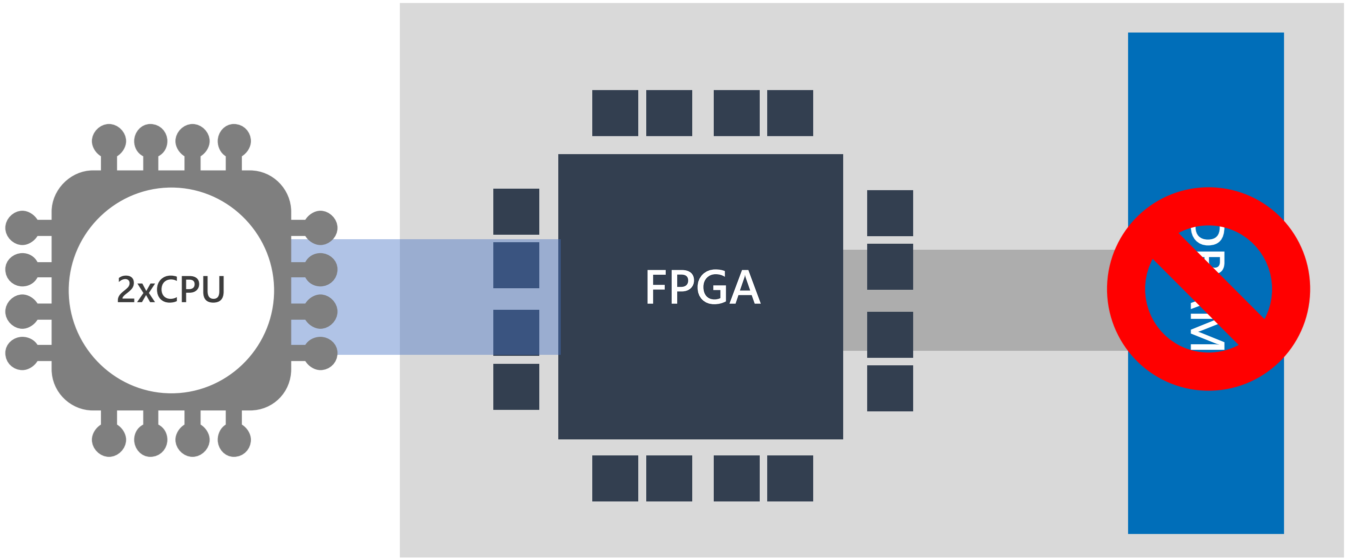




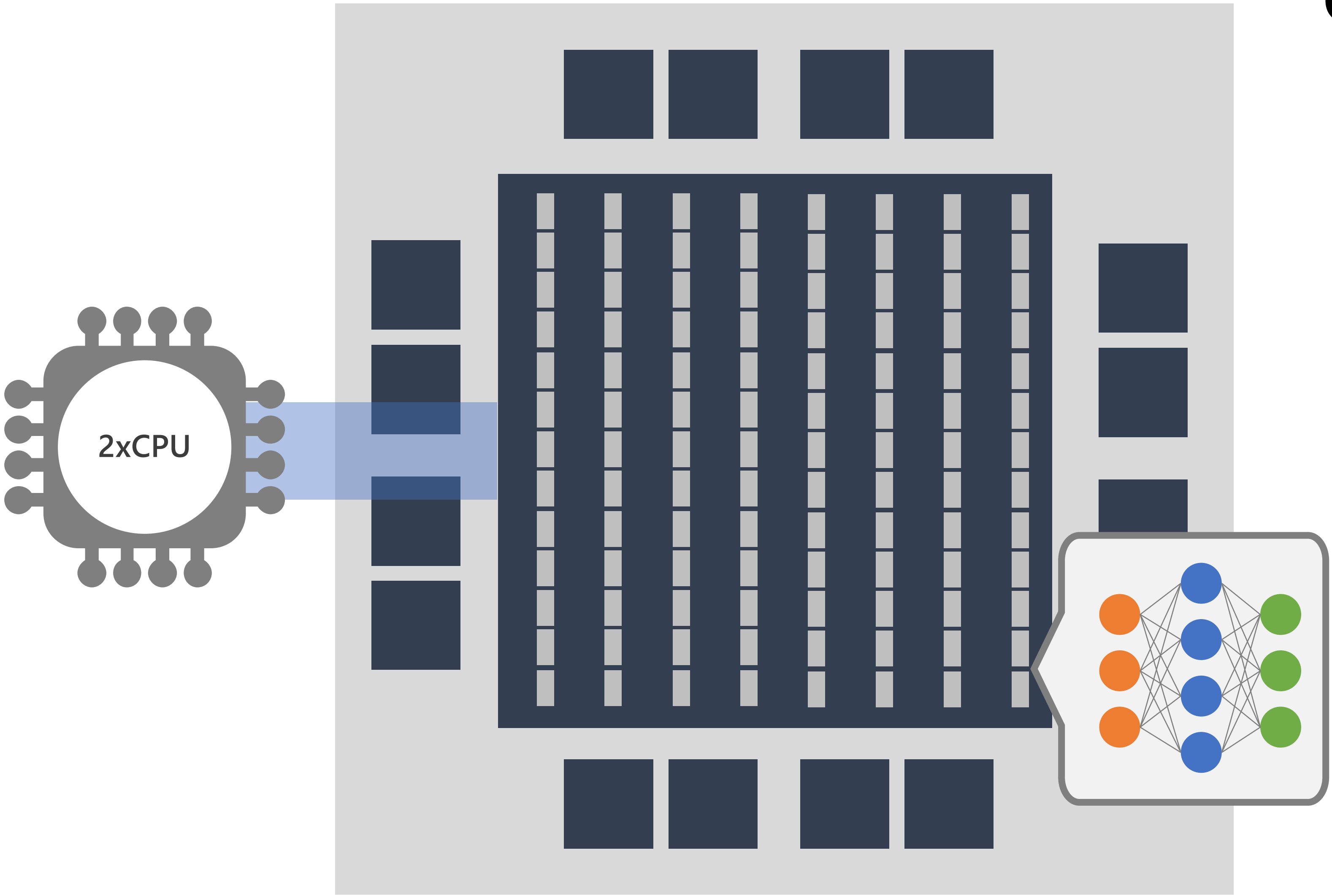
Batching improves HW utilization but increases latency

Ideally want high HW utilization at low batch sizes

Alternative: “Persistent” Neural Nets



Alternative: "Persistent" Neural Nets



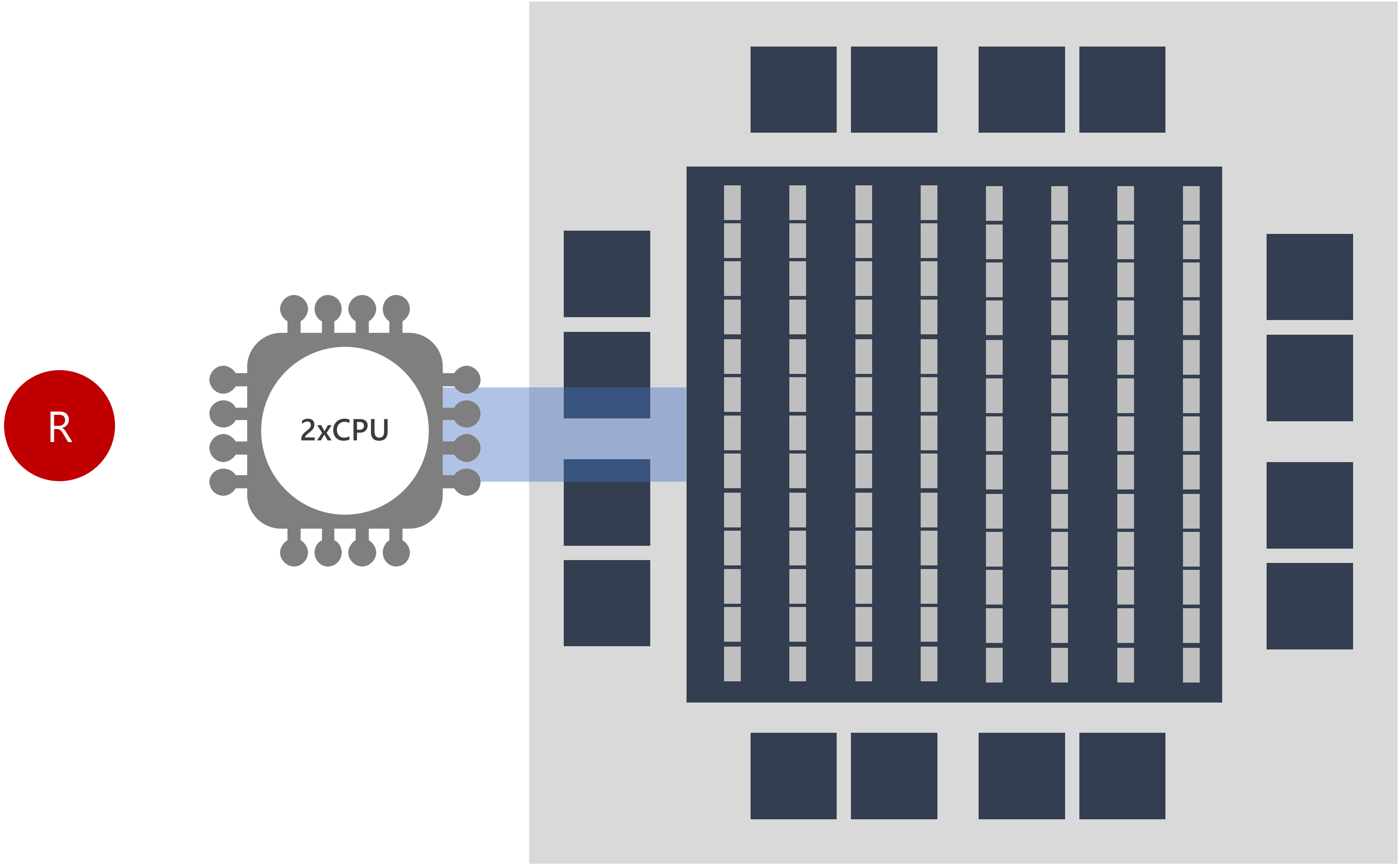
Observations

State-of-art FPGAs have O(10K) distributed Block RAMs O(10MB)
➔ Tens of TB/sec of memory BW

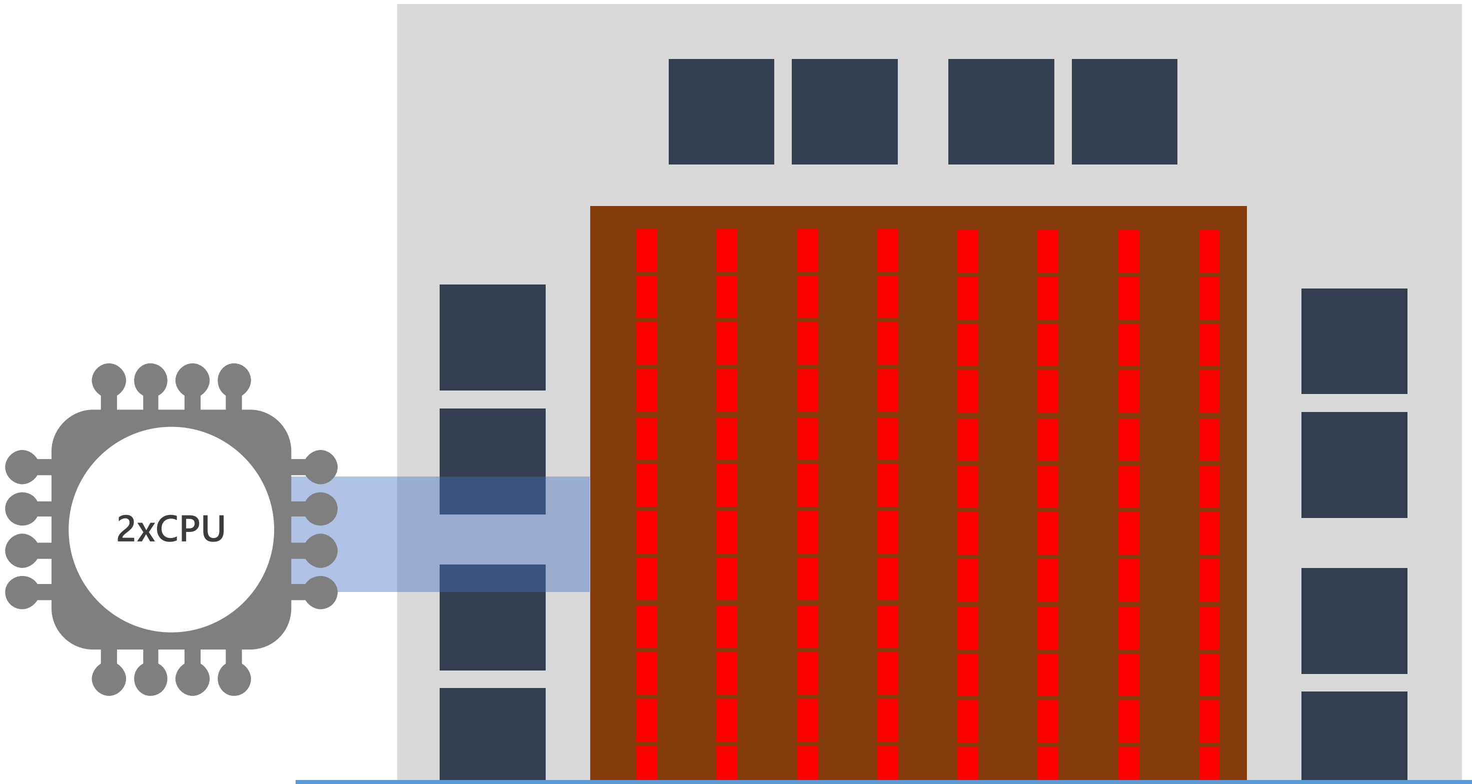
Large-scale cloud services and DNN models run persistently

Solution: persist all model parameters in FPGA on-chip memory during service lifetime

Alternative: "Persistent" Neural Nets



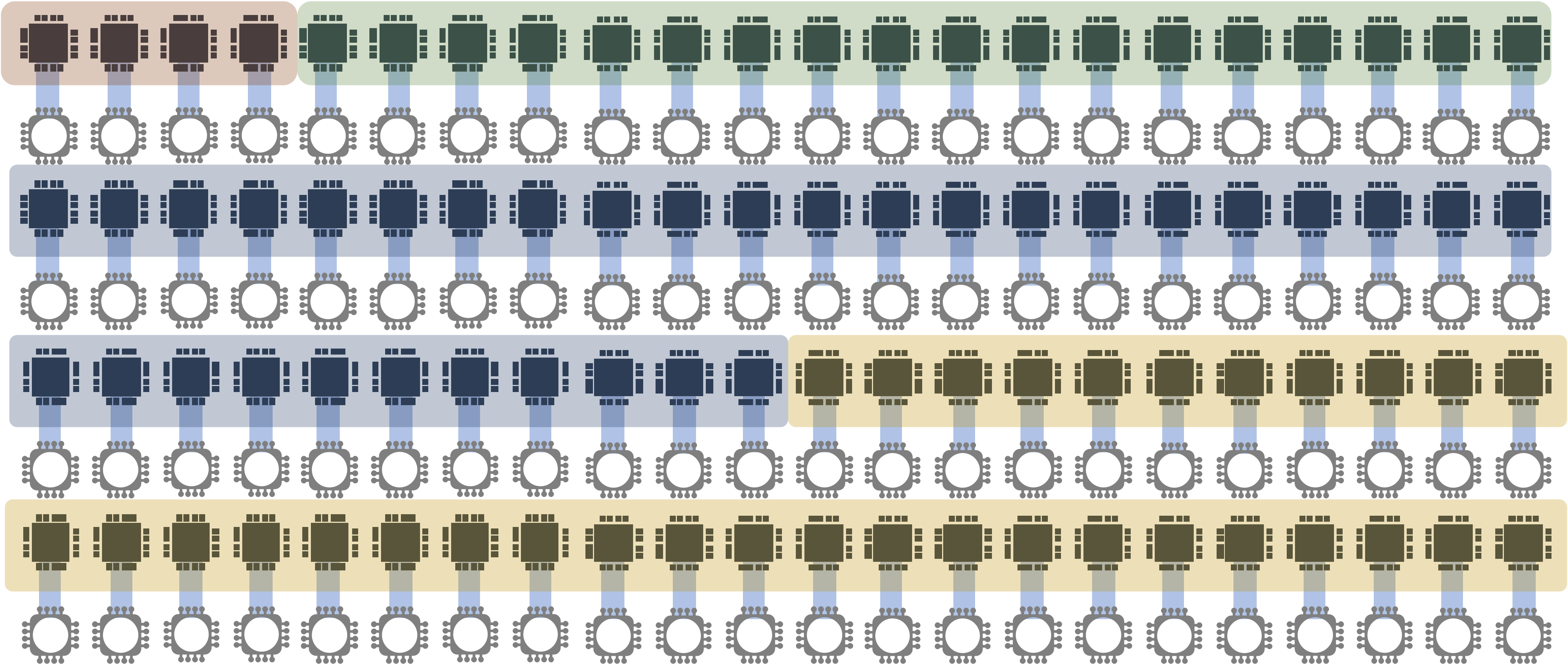
Alternative: “Persistent” Neural Nets



When single request arrives, all chip resources (on-chip memories and compute units) are used to process a single query (no batching required)

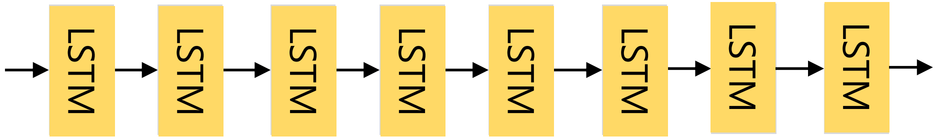
What if model doesn’t fit in single FPGA?

Solution: Persistency at Datacenter Scale

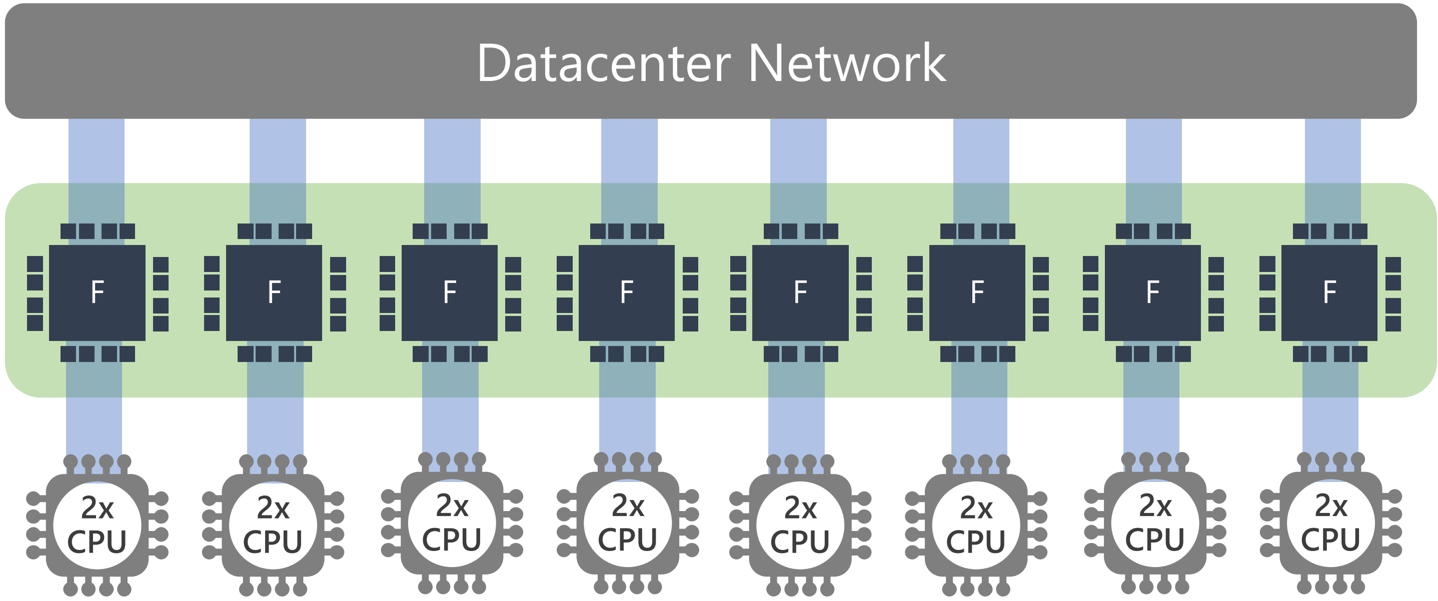


Multiple FPGAs at datacenter scale can form a persistent DNN HW microservice, enabling scale-out of models at ultra-low latencies

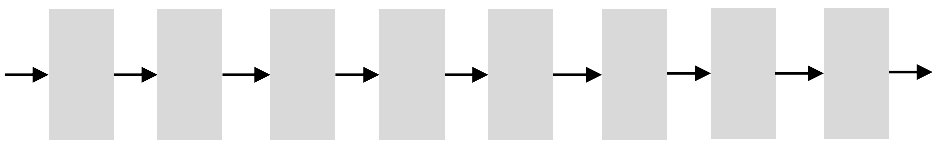
Inter-Layer Pipeline Parallelism

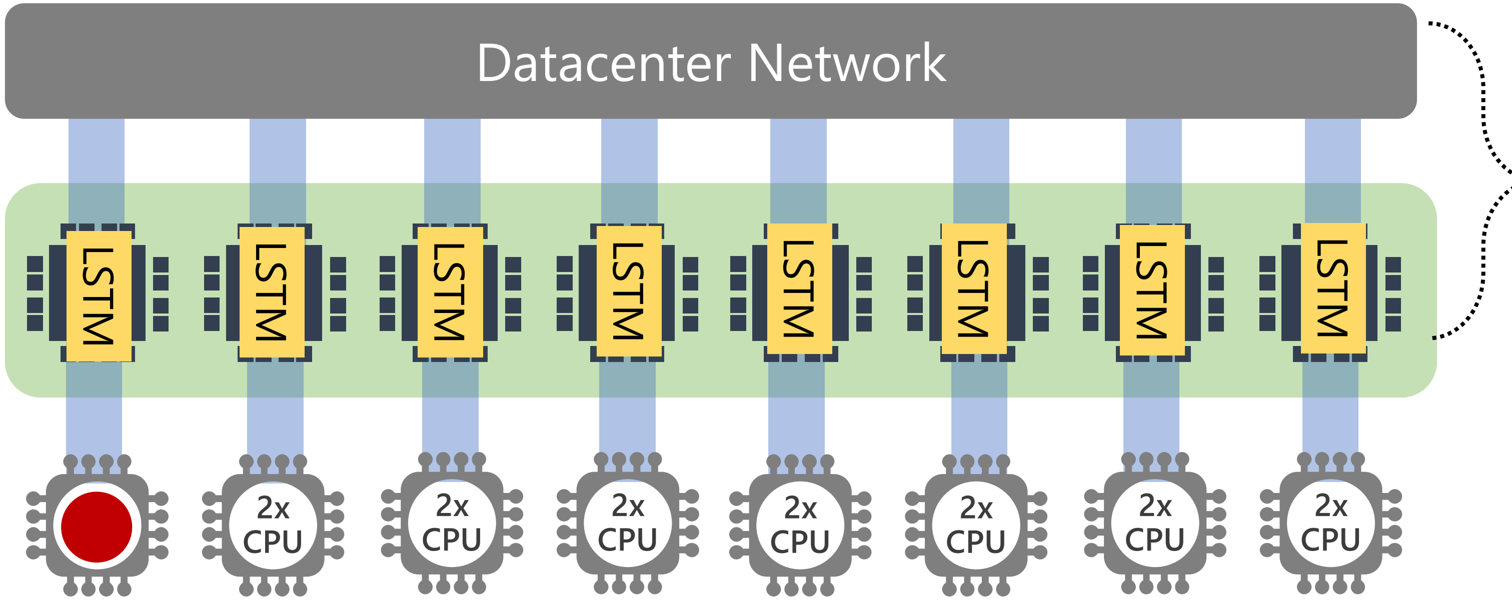


$$
\begin{aligned}
f_t &= \sigma_g(\boxed{W_f x_t + U_f h_{t-1}} + b_f) \\
i_t &= \sigma_g(\boxed{W_i x_t + U_i h_{t-1}} + b_i) \\
o_t &= \sigma_g(\boxed{W_o x_t + U_o h_{t-1}} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(\boxed{W_c x_t + U_c h_{t-1}} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$



Inter-Layer Pipeline Parallelism

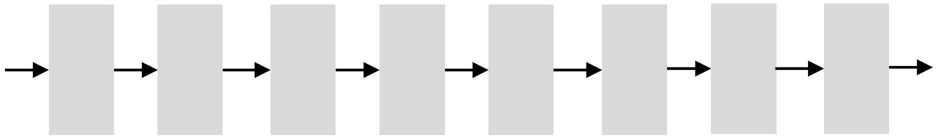


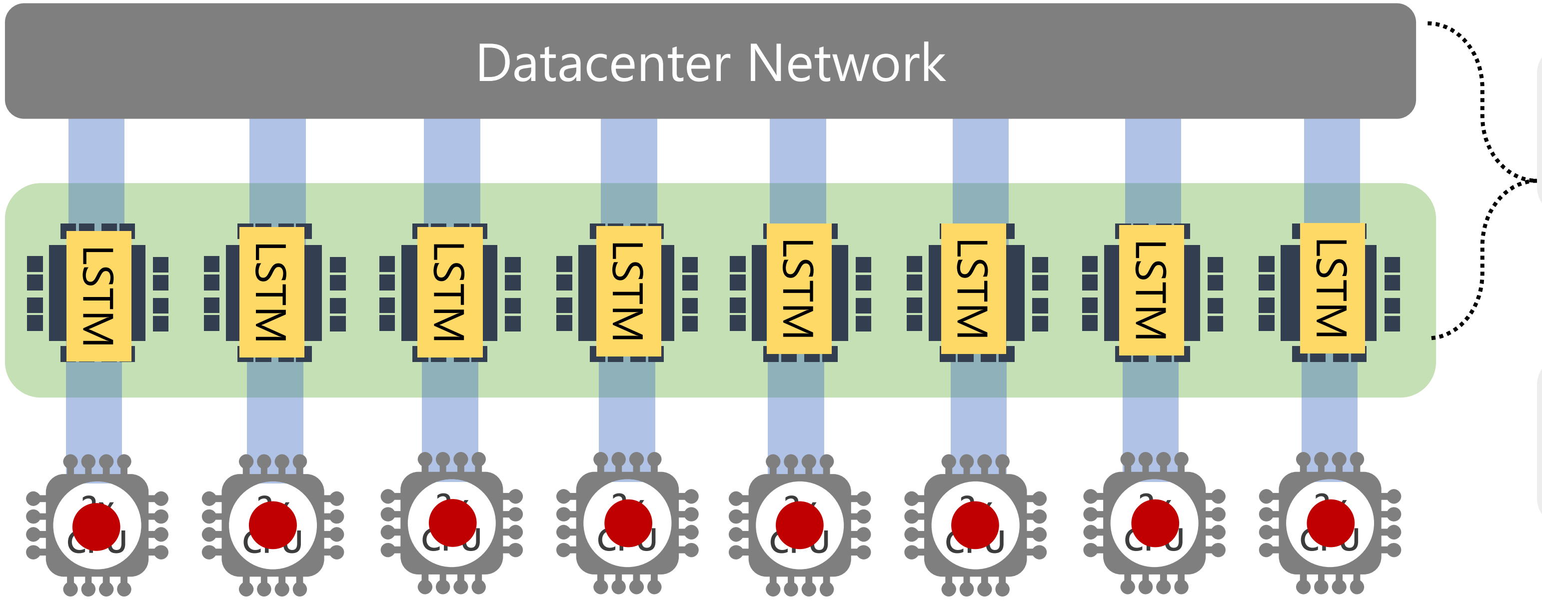


FPGAs communicate directly < 2us/hop

DNN HWMS shared by all CPUs

Inter-Layer Pipeline Parallelism

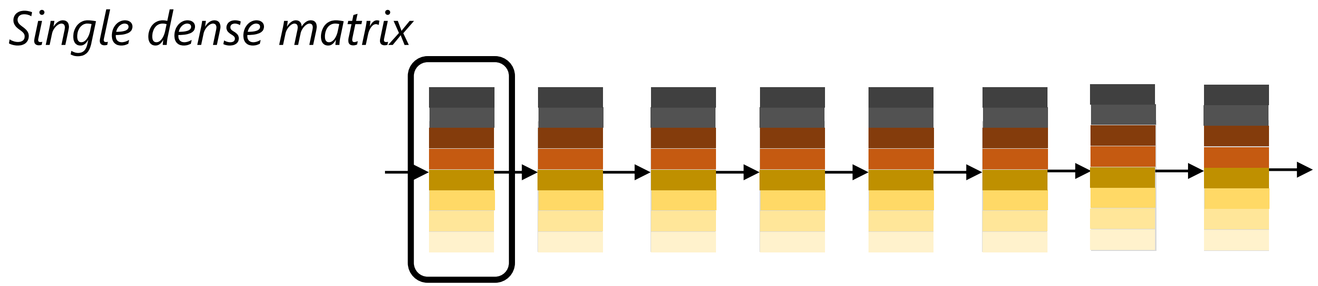


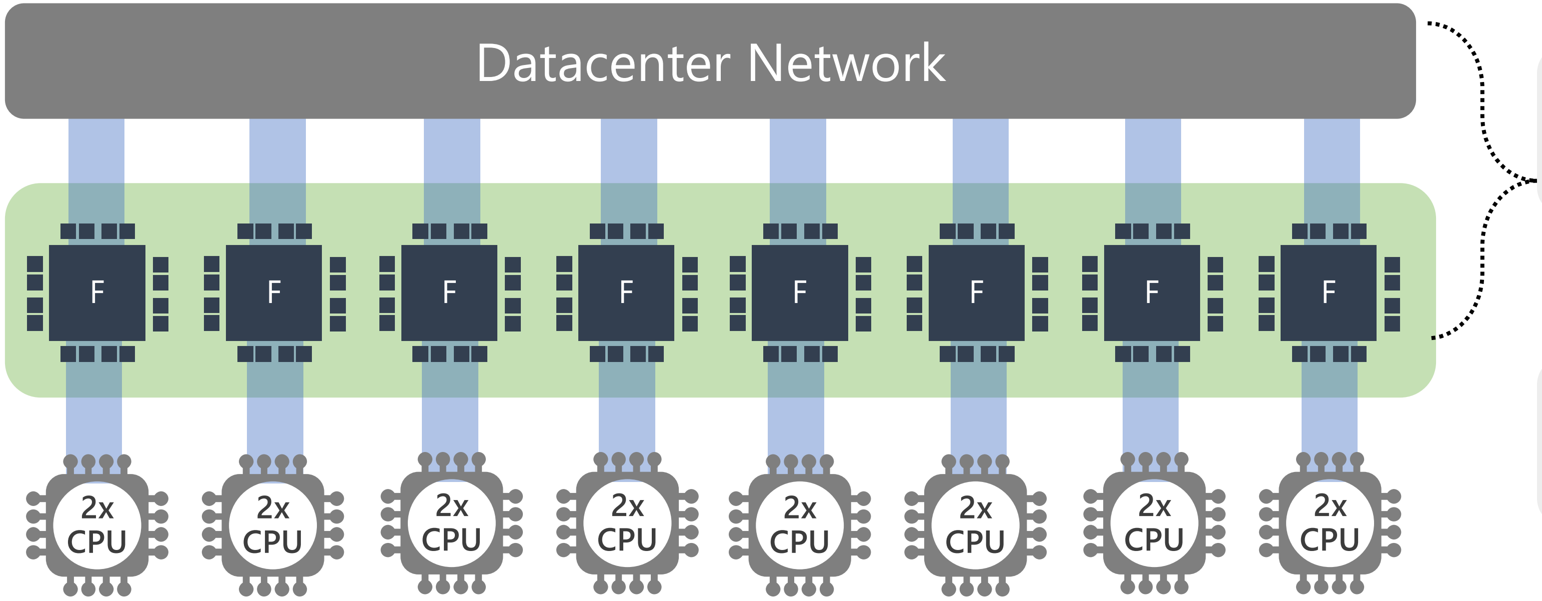


FPGAs communicate directly < 2us/hop

DNN HWMS shared by all CPUs

Intra-Layer Parallelism

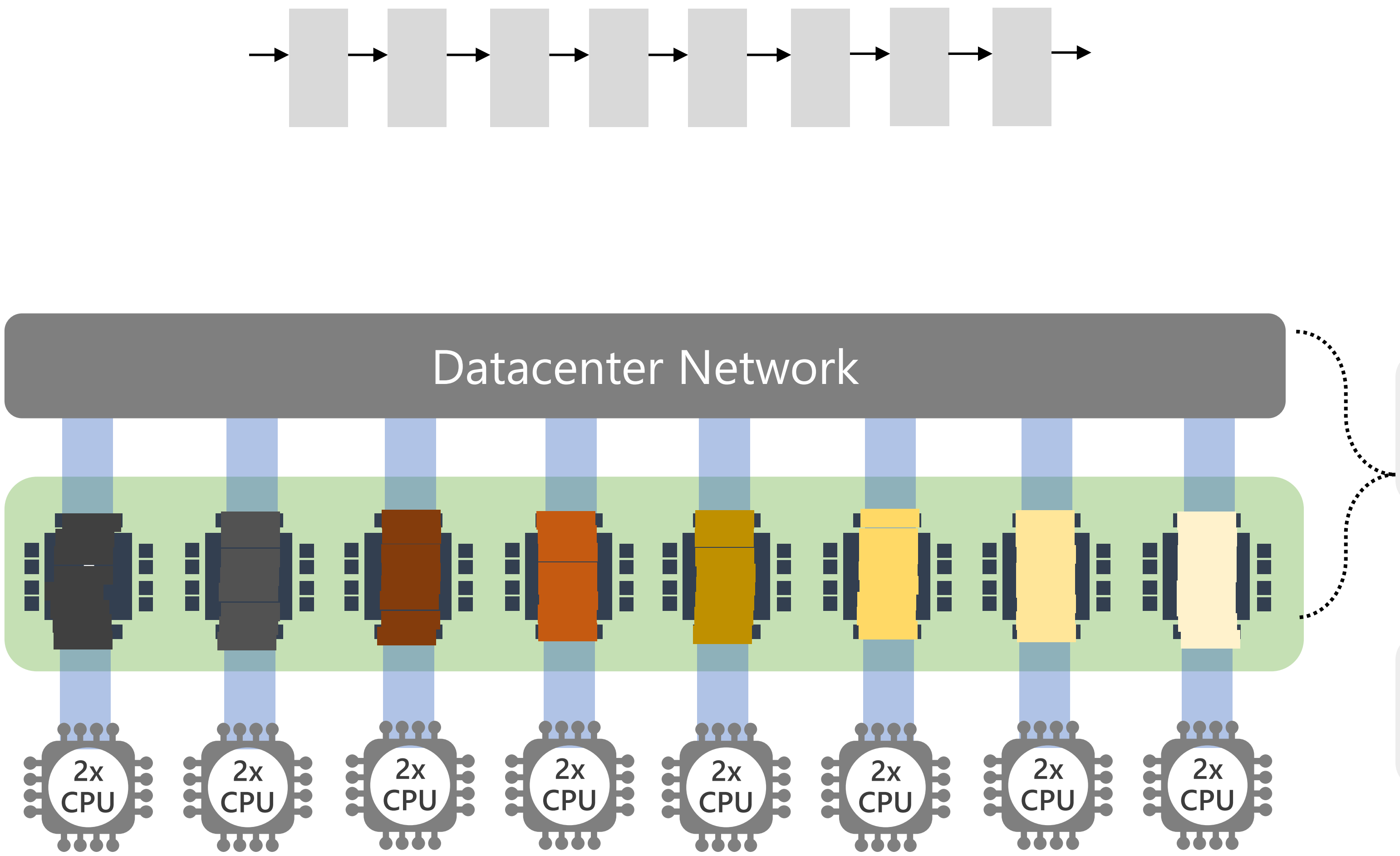




FPGAs communicate directly < 2us/hop

DNN HWMS shared by all CPUs

Intra-Layer Parallelism



FPGAs communicate directly < 2us/hop

DNN HWMS shared by all CPUs

The BrainWave Stack

Compiler & Runtime

Architecture

Microarchitecture

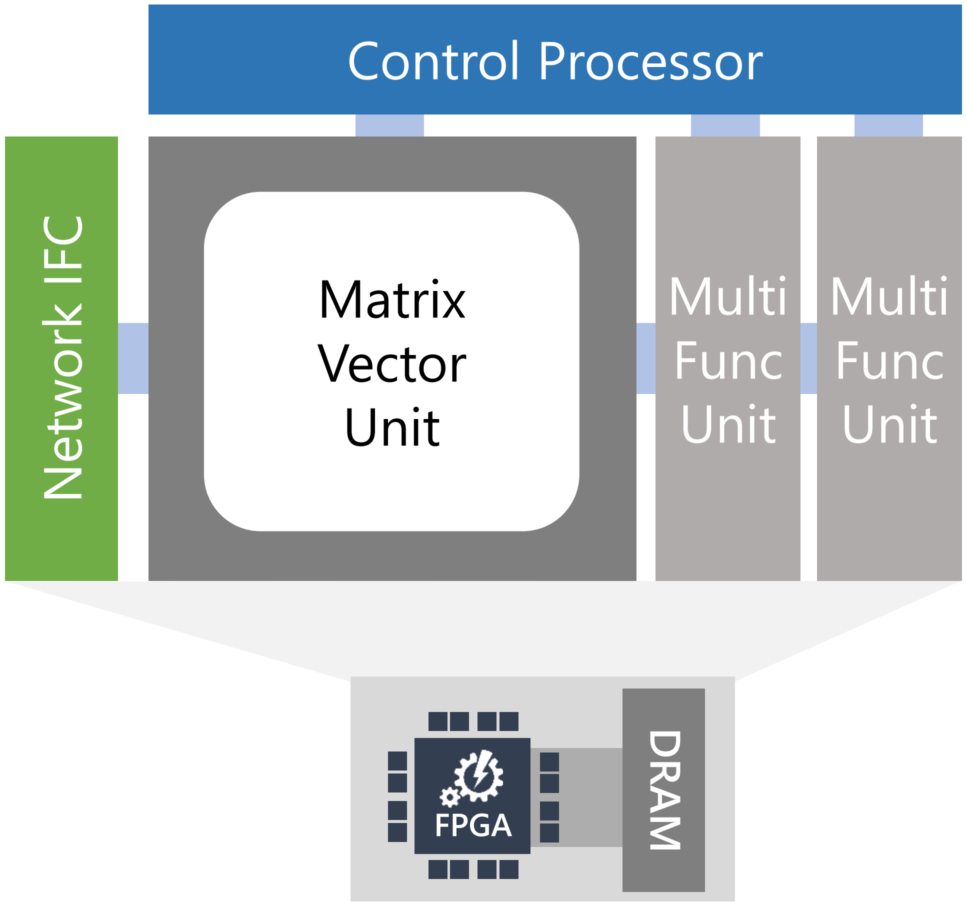
Persistency at Scale

HW Microservices
on Intel FPGAs

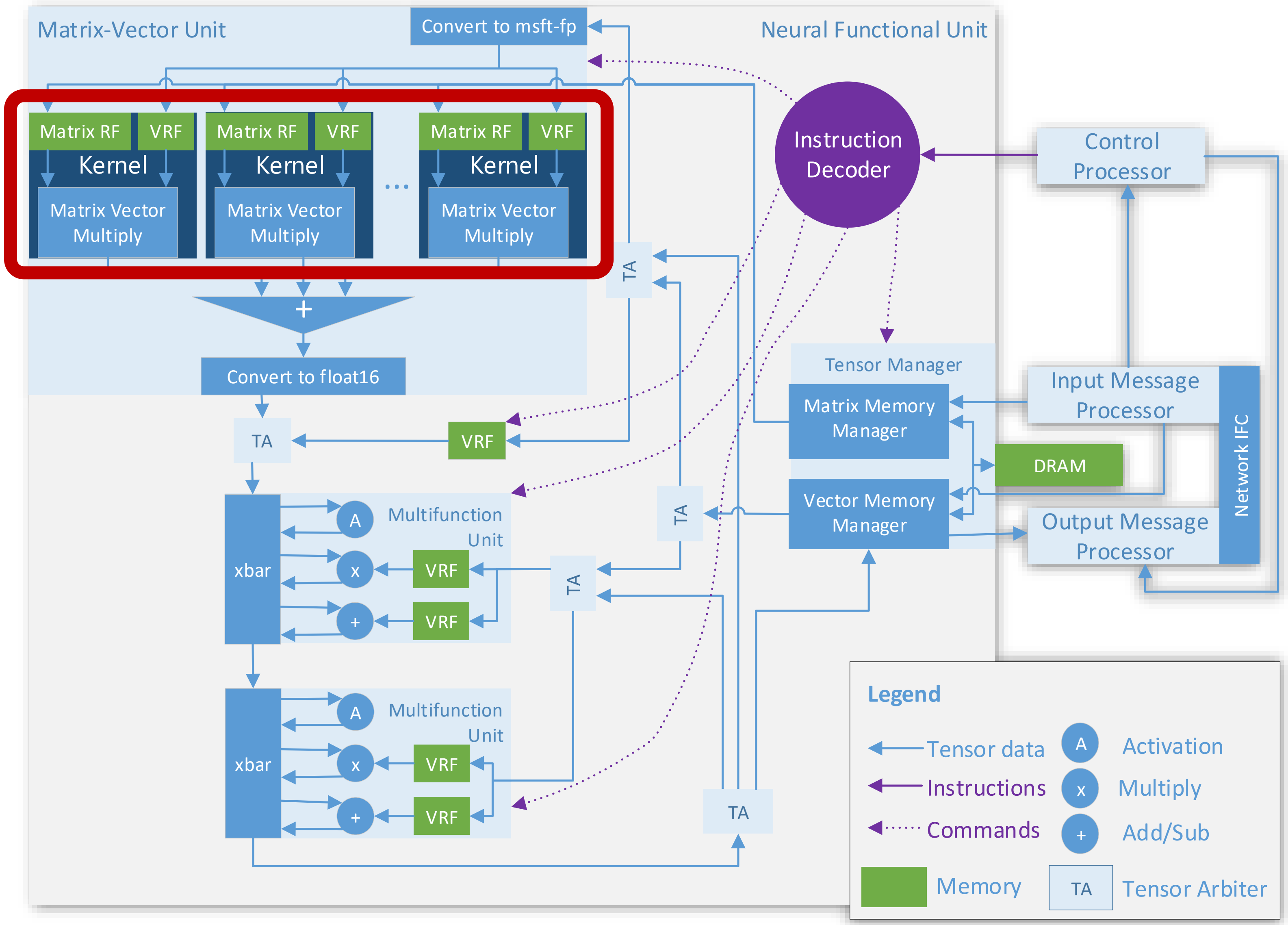
BrainWave Soft DPU Architecture

Core Features

- Single-threaded C programming model (no RTL)
- ISA with specialized instructions: dense matmul, convolutions, non-linear activations, vector operations, embeddings
- Proprietary parameterizable narrow precision format wrapped in float16 interfaces
- Parameterizable microarchitecture and scalable to large FPGAs (~1M ALMs)
- Fully integrated with HW microservices (network-attached)
- P2P protocol to CPU hosts and FPGAs
- Easy to extend ISA with custom operators



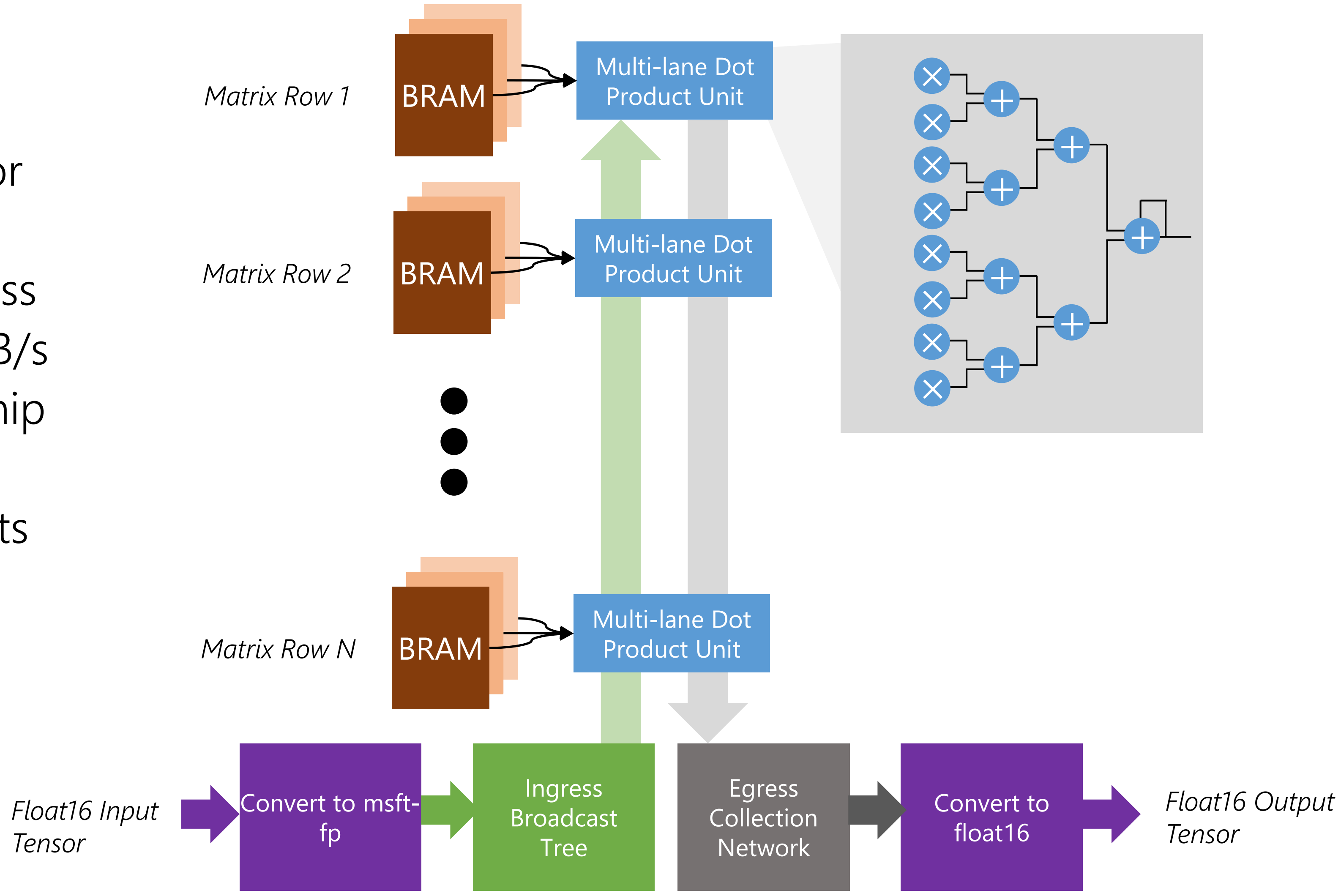
BrainWave Soft DPU Microarchitecture



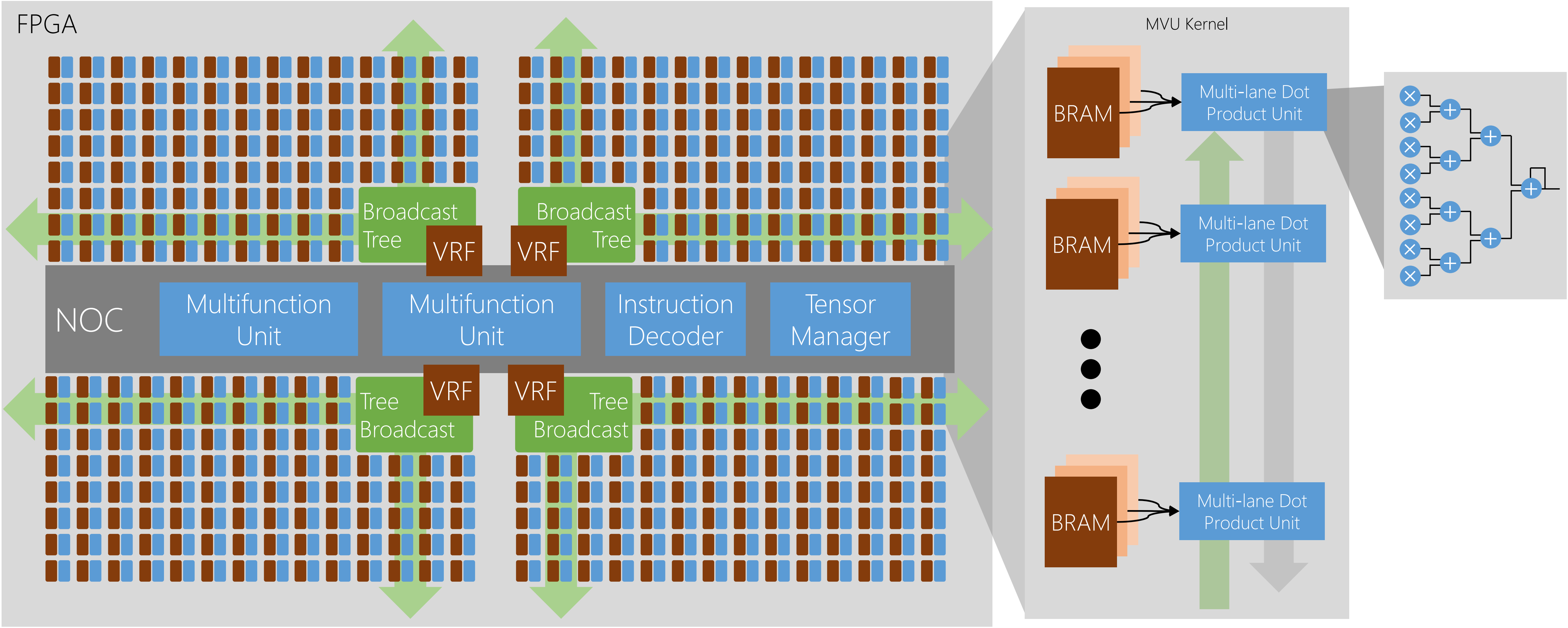
Matrix Vector Unit

Features

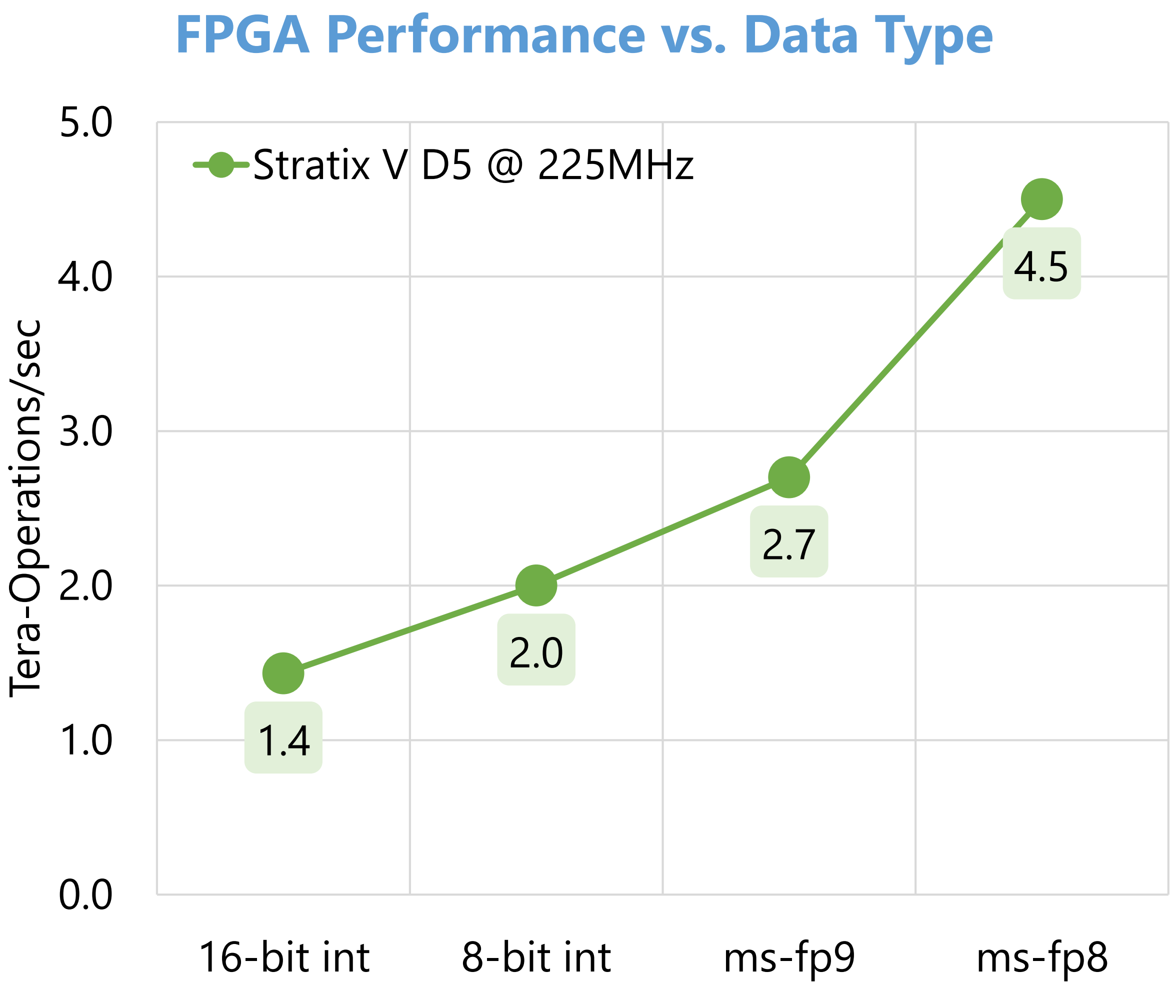
- Optimized for batch 1 matrix-vector multiplication
- Matrices distributed row-wise across 1K-10K banks of BRAM, up to 20 TB/s
- Can scale to use all available on-chip BRAMs, DSPs, and soft logic
- In-situ conversion of float16 weights and activations to internal format
- Dense dot product units map efficiently to soft logic and DSPs



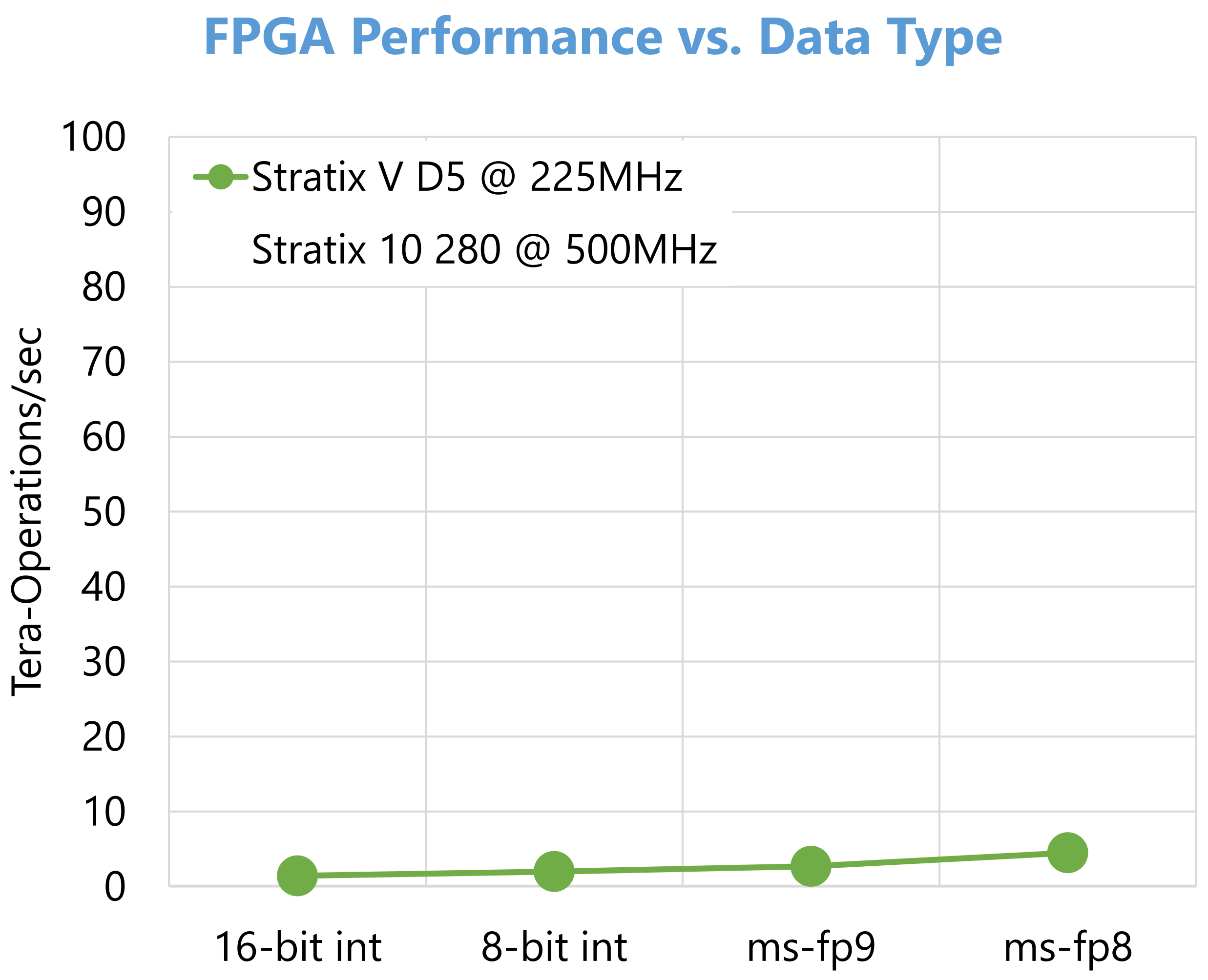
Matrix Vector Unit



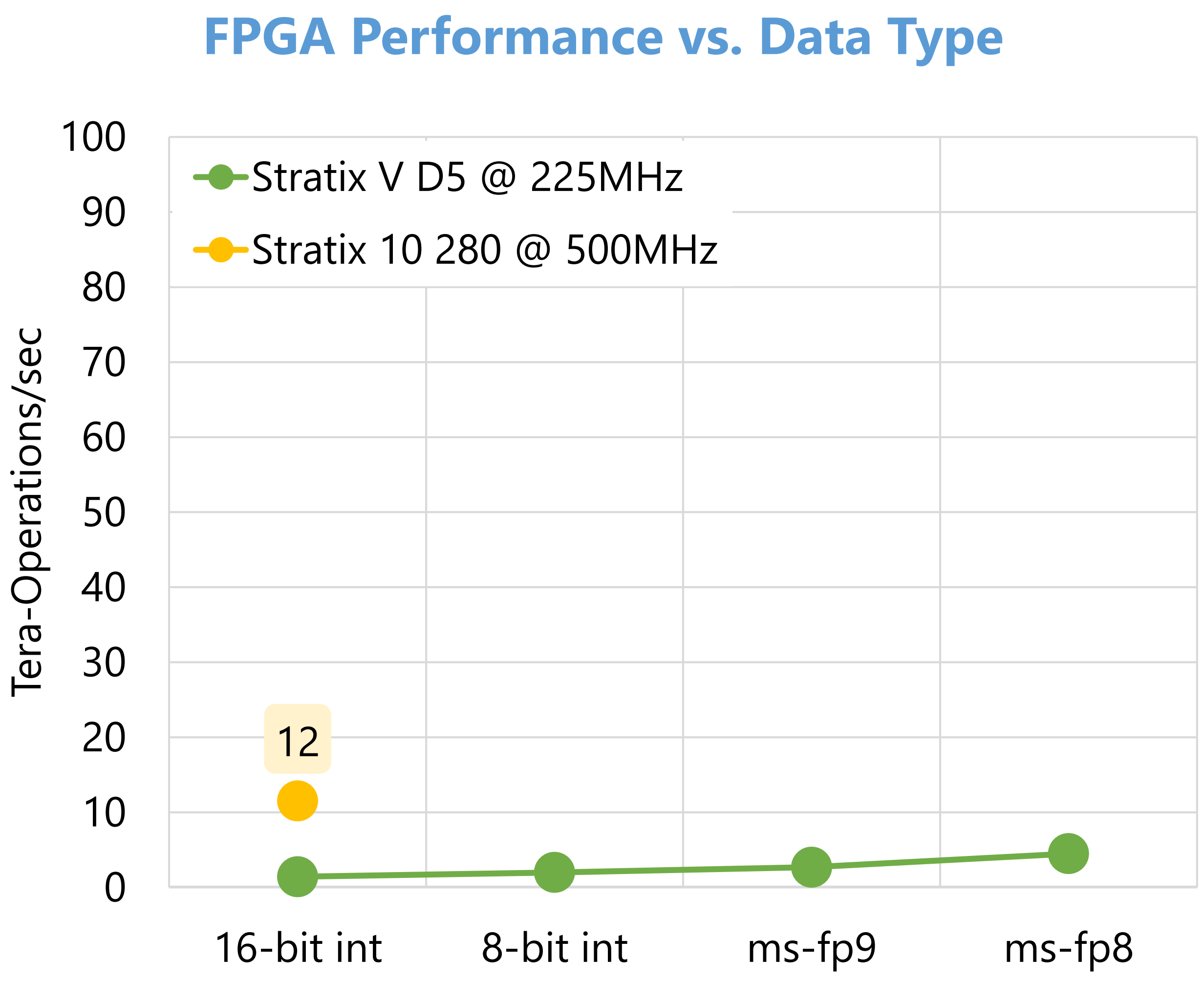
Narrow Precision Inference on FPGAs



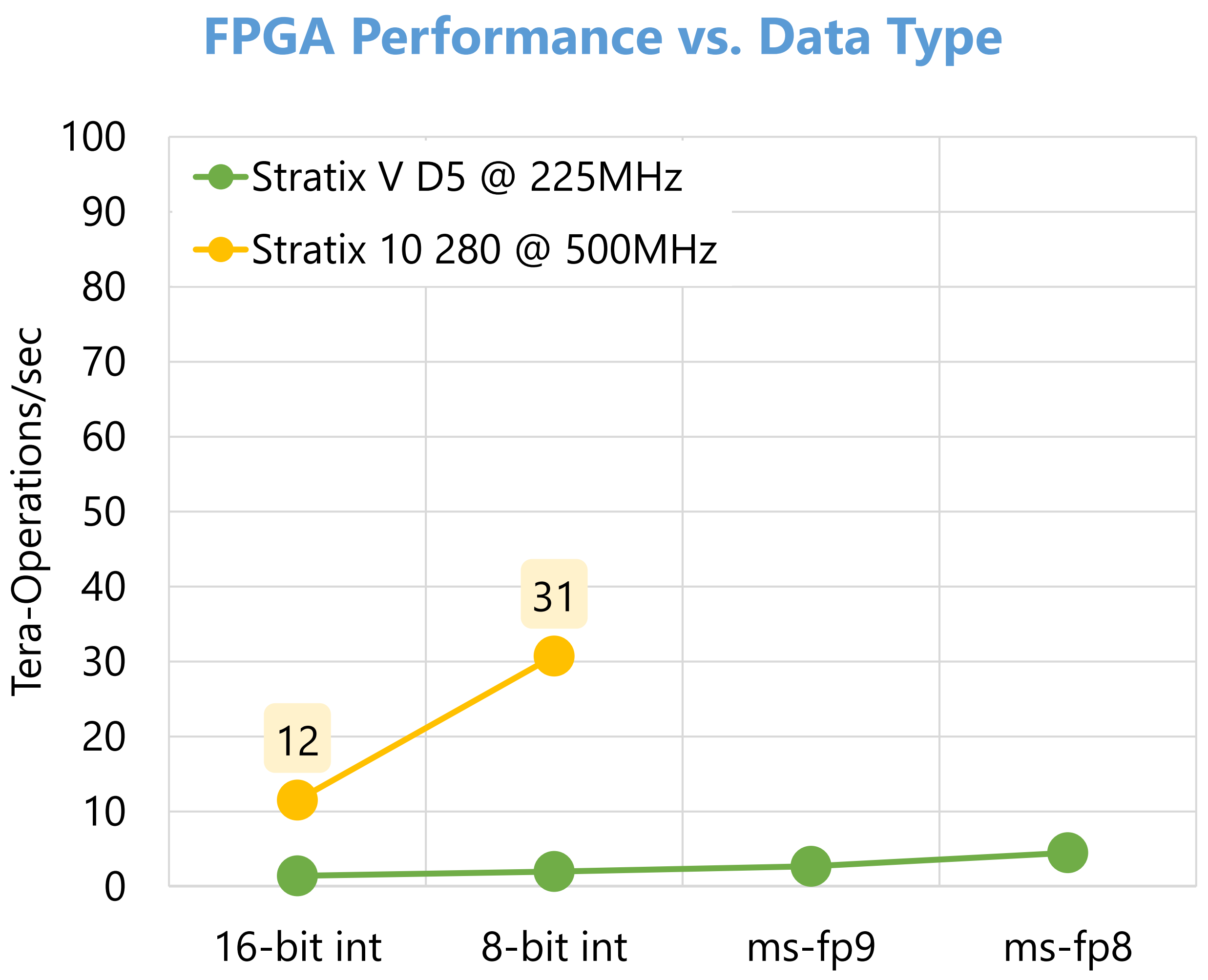
Narrow Precision Inference on FPGAs



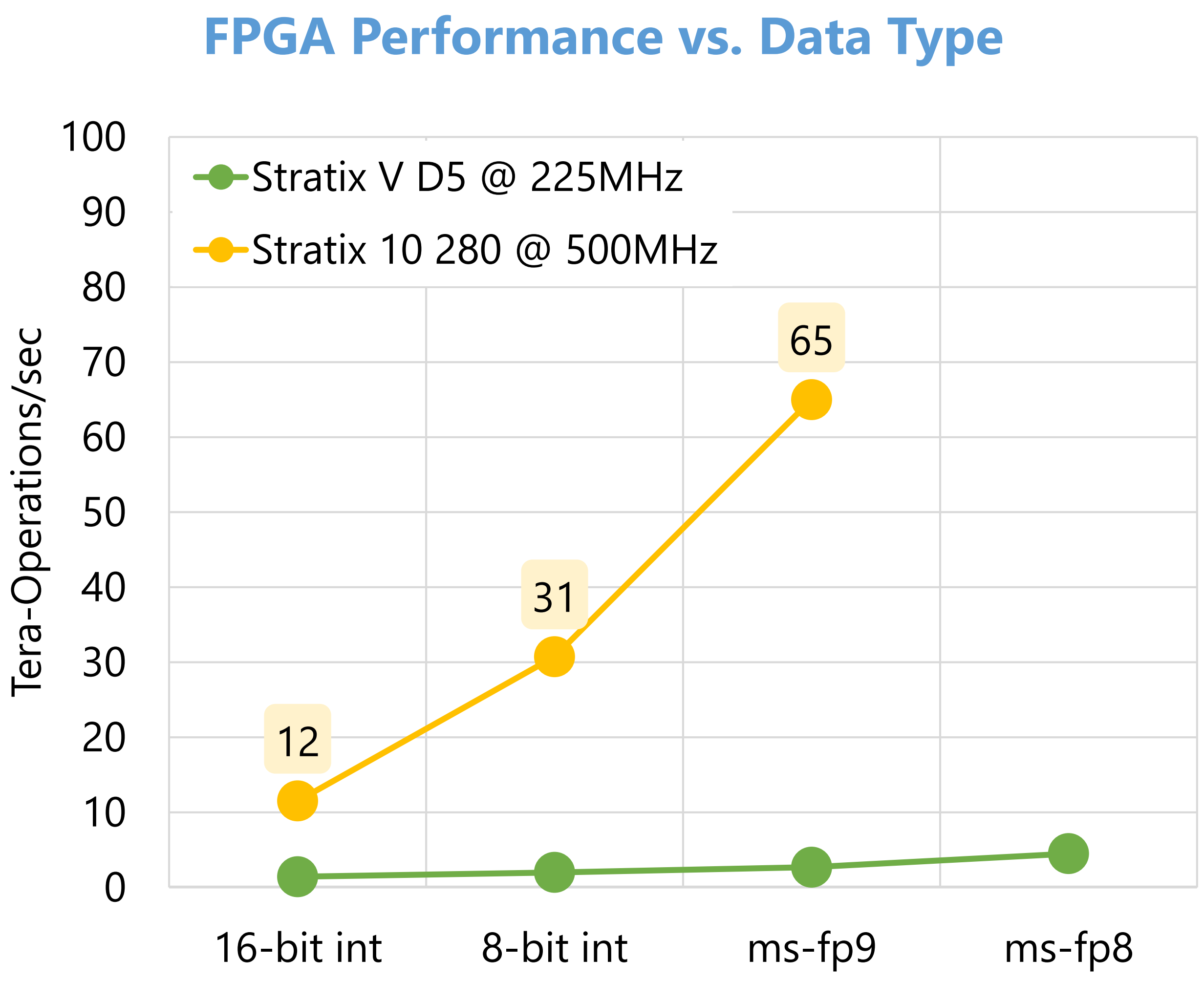
Narrow Precision Inference on FPGAs



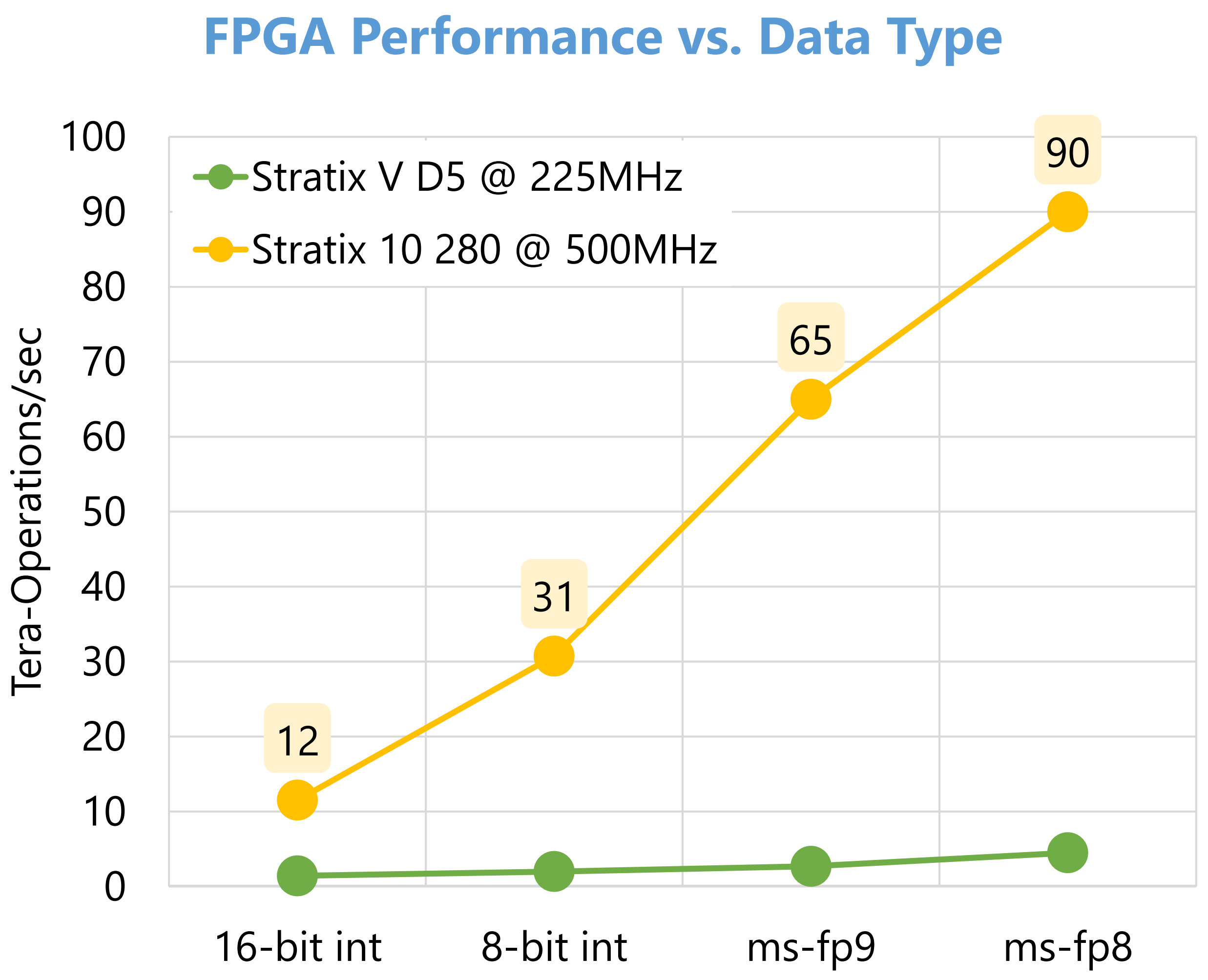
Narrow Precision Inference on FPGAs



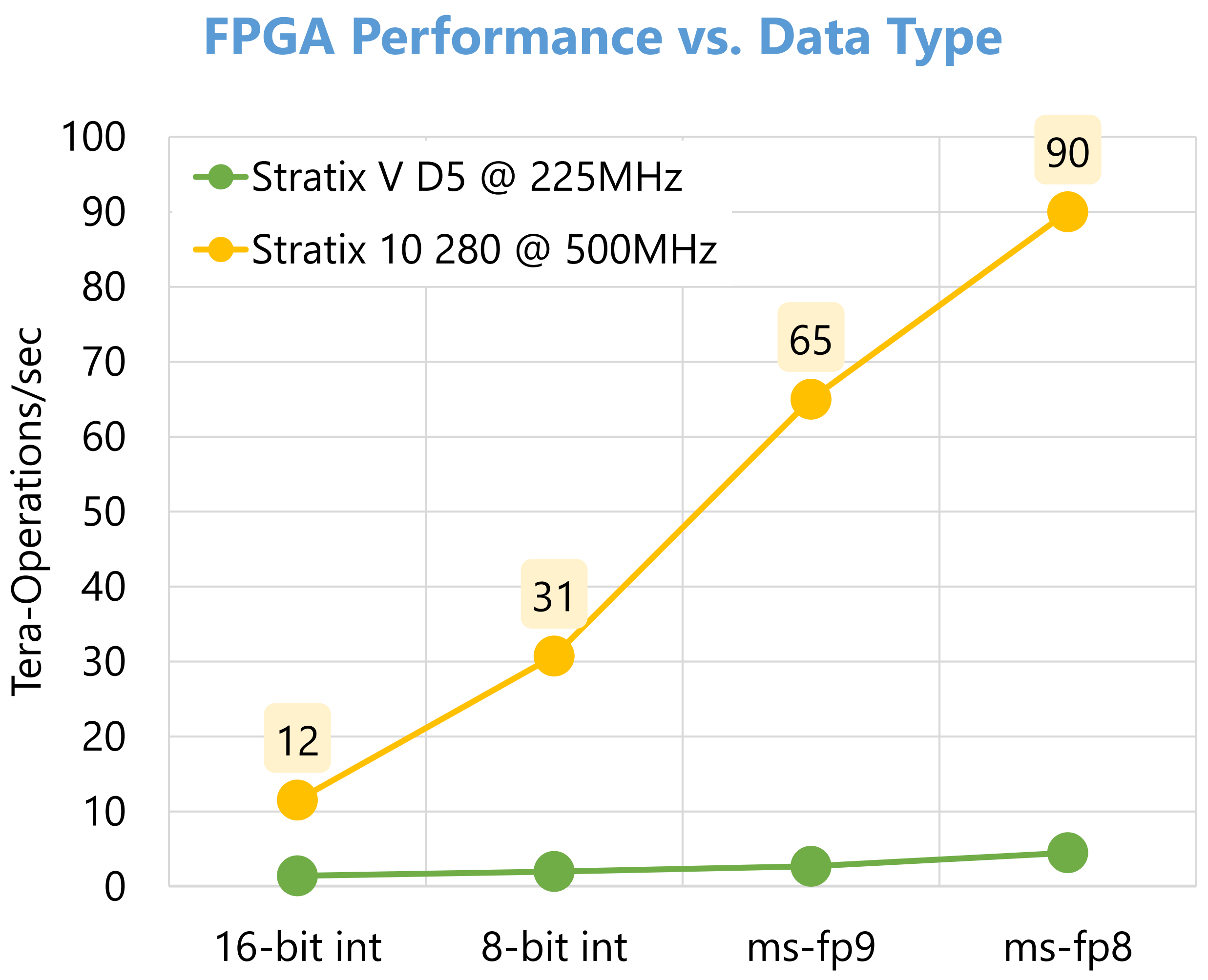
Narrow Precision Inference on FPGAs

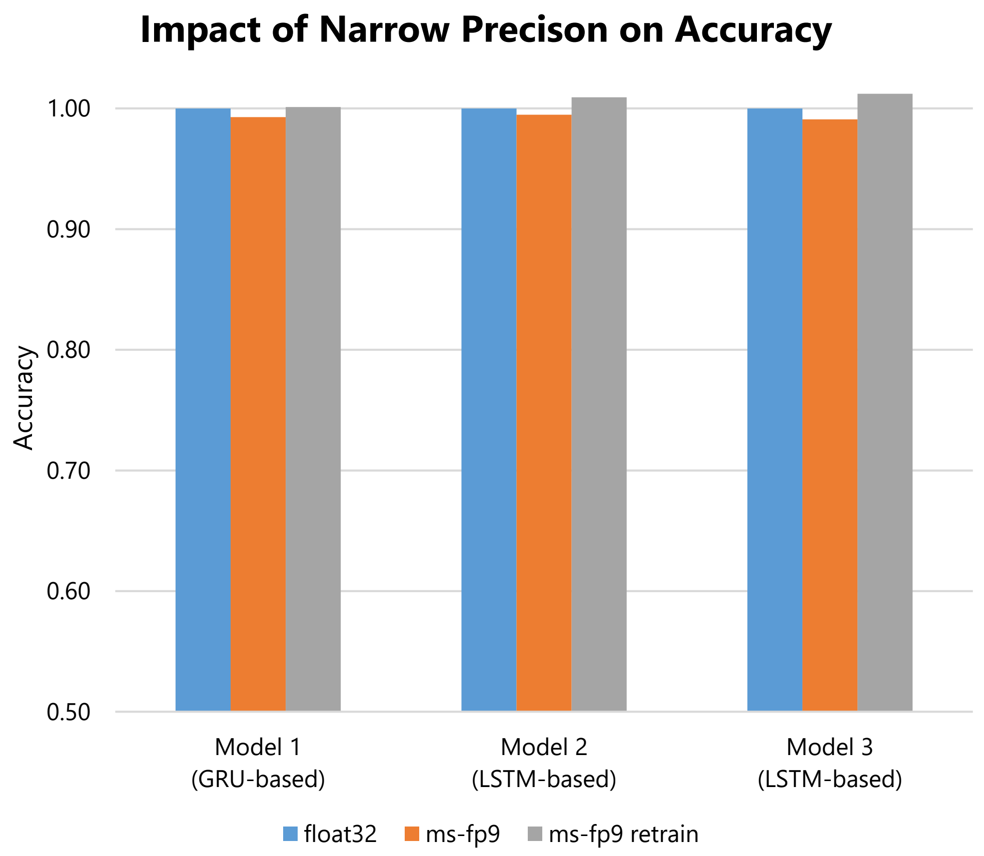


Narrow Precision Inference on FPGAs

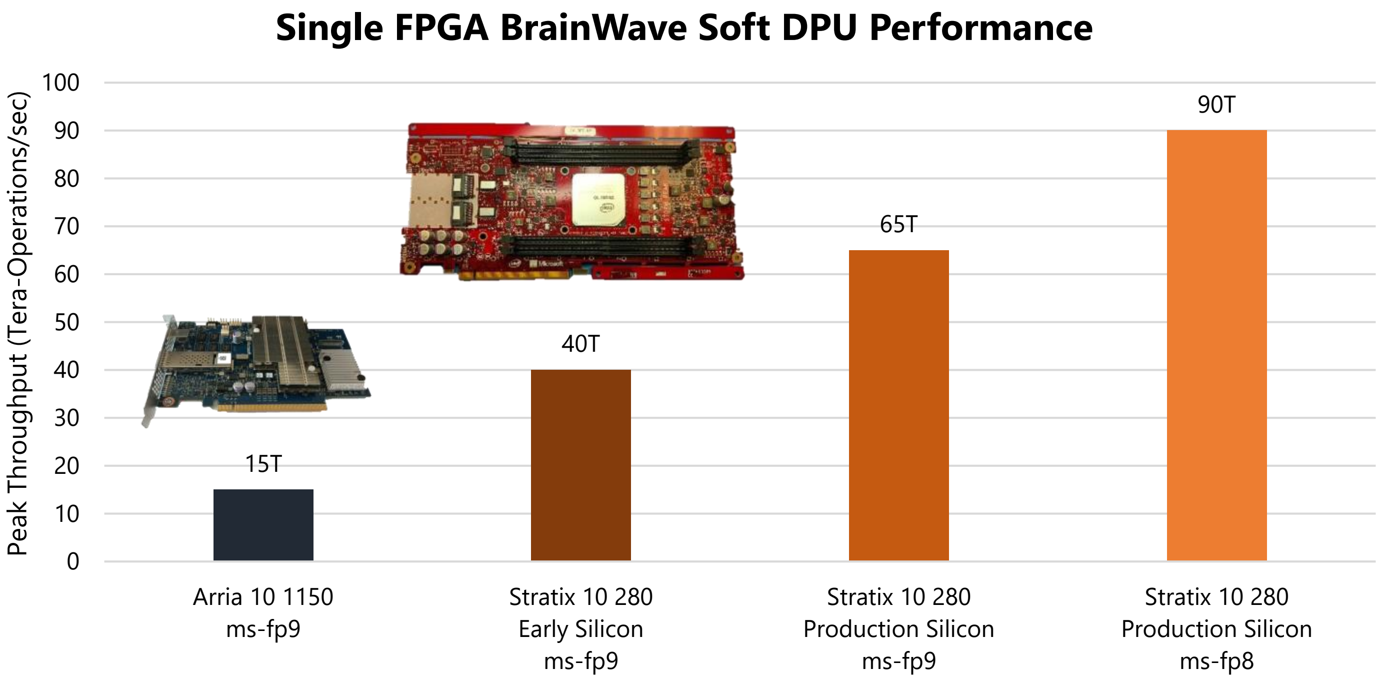


Narrow Precision Inference on FPGAs

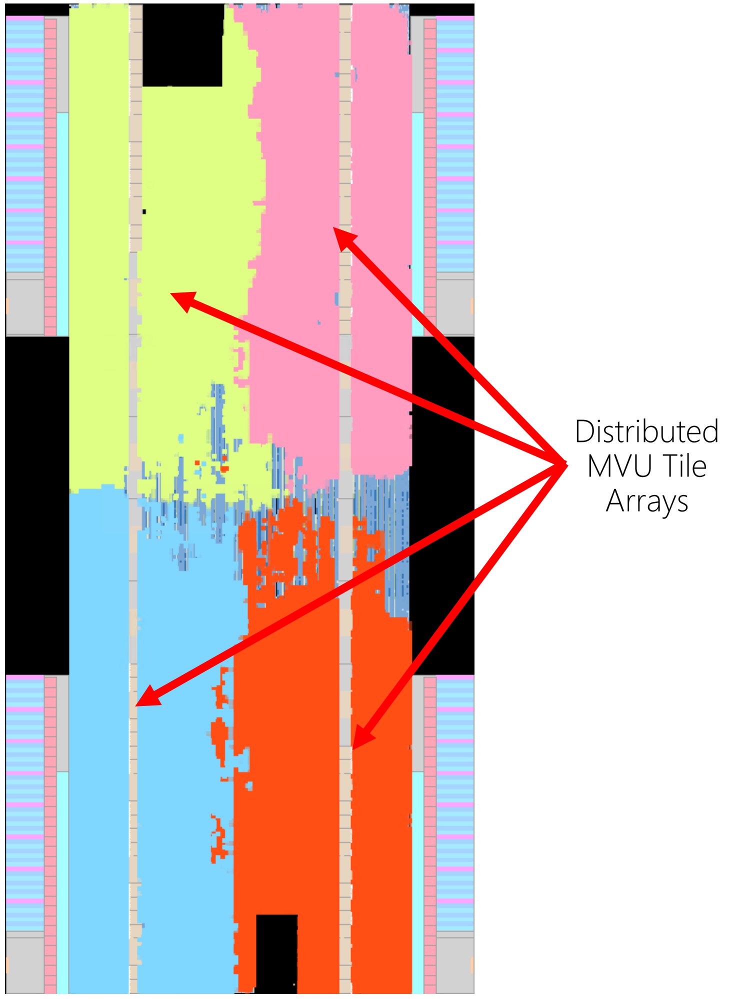




BrainWave Soft DPU Performance



Arria 10 1150 (20nm)	Stratix 10 280 Early Silicon (14nm)
ms-fp9	ms-fp9
316K ALMs (74%)	858K ALMs (92%)
1442 DSPs (95%)	5,760 DSPs (100%)
2,564 M20Ks (95%)	8,151 M20Ks (70%)
160 GOPS/W	320 GOPS/W → 720 GOPS/W (production)



BrainWave Soft DPU
Floorplan on Stratix 10 280

Conclusion

Microsoft BrainWave is a powerful platform for an accelerated AI cloud

- Runs on Microsoft's hyperscale infrastructure with FPGAs
- Achieves excellent performance at low batch sizes via persistency and narrow precision
- Adaptable to precision and changes in future AI algorithms

BrainWave running on Hardware Microservices will push the boundary of what is possible to deploy in the cloud

- Deeper/larger CNNs for more accurate computer vision
- Higher dimensional RNNs toward human-like natural language processing
- State-of-the-art speech
- And much more…

Stay tuned for announcements about external availability.

Thank you!